

Как при помощи бумаги, карандаша и алгоритма Raft достичь консенсуса

Ярослав Дынников

Picodata



Слайды: <https://rosik.github.io/2023-highload>

О чем речь

Кластер — это группа процессов, работающих совместно и представляющихся пользователю единым компьютерным ресурсом.

О чем речь

Кластер — это группа процессов, работающих совместно и представляющихся пользователю единым компьютерным ресурсом.

Задача

- Есть несколько серверов.
- Надо достичь консенсуса.

О чем речь

Кластер — это группа процессов, работающих совместно и представляющихся пользователю единым компьютерным ресурсом.

Задача

- Есть несколько серверов.
- Надо достичь консенсуса.
- В ненадежной сети.

О чем речь

Кластер — это группа процессов, работающих совместно и представляющихся пользователю единым компьютерным ресурсом.

Задача

- Есть несколько серверов.
- Надо достичь консенсуса.
- В ненадежной сети.

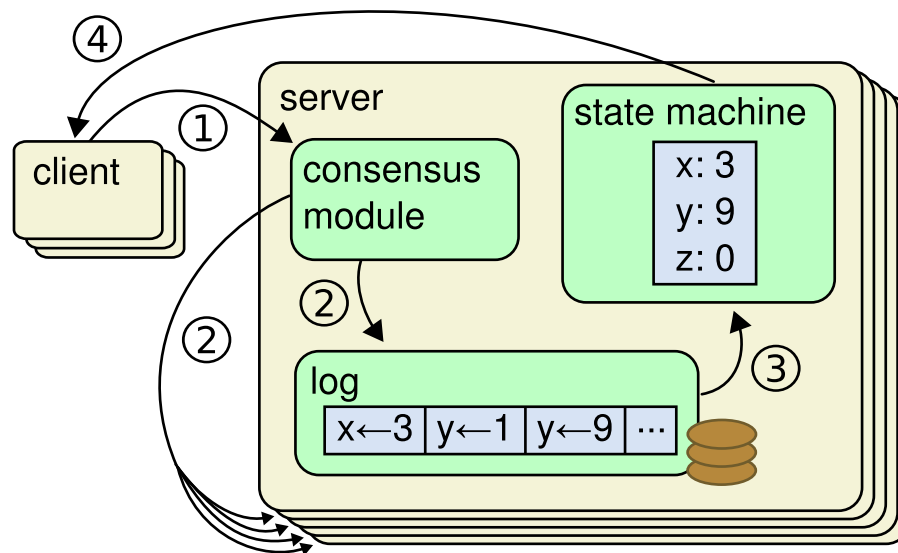
Решение — Raft

- In search of Understandable Consensus Algorithm.
- Diego Ongaro and John Ousterhout. Stanford University.
- <https://raft.github.io>

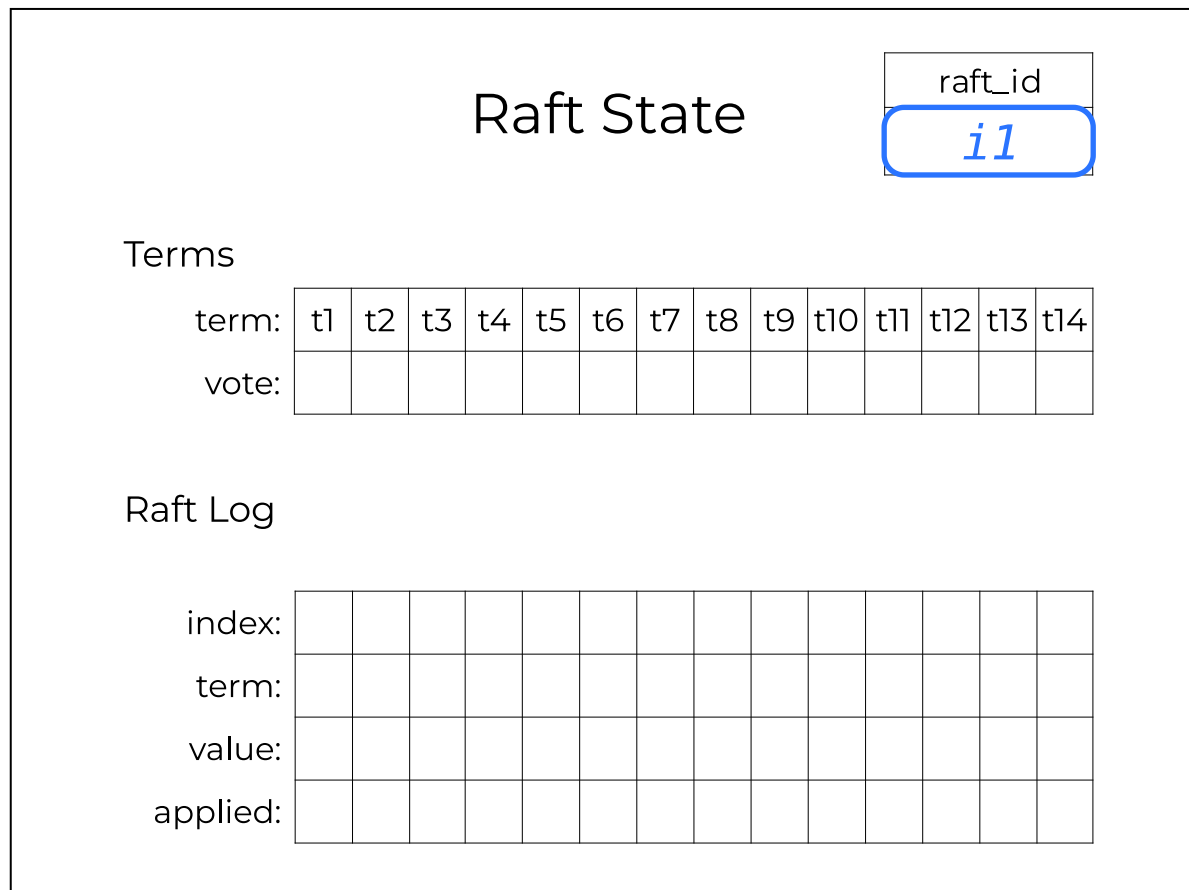
Raft



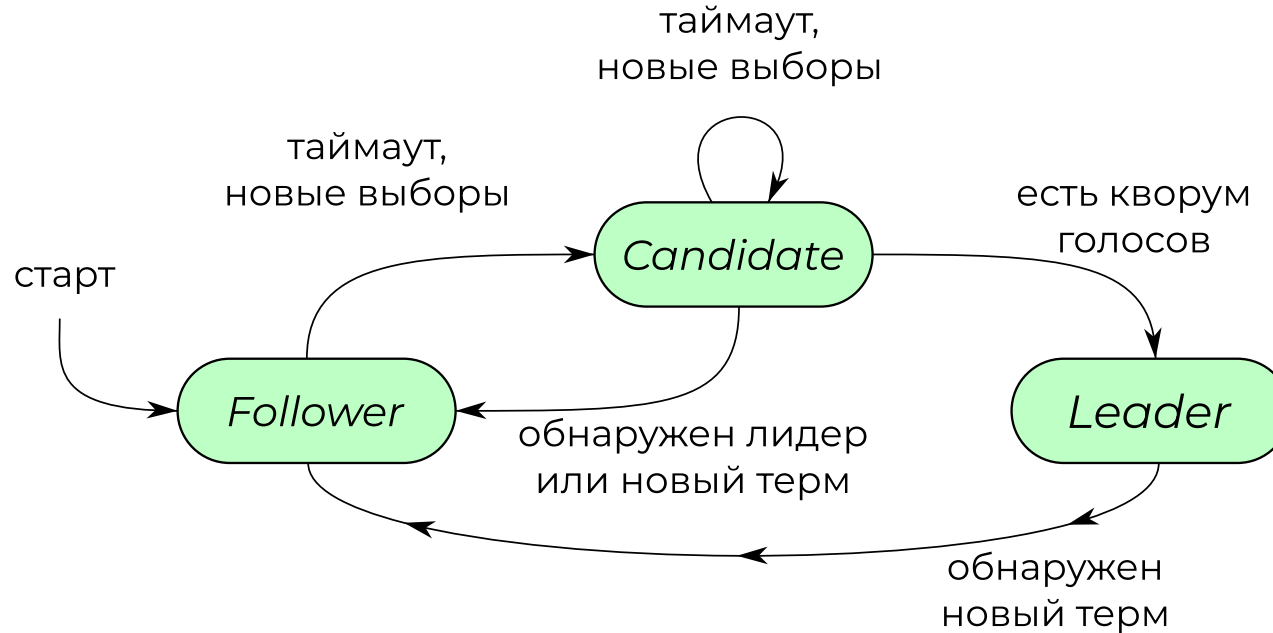
Реплицируемый конечный автомат



Персистентное хранилище



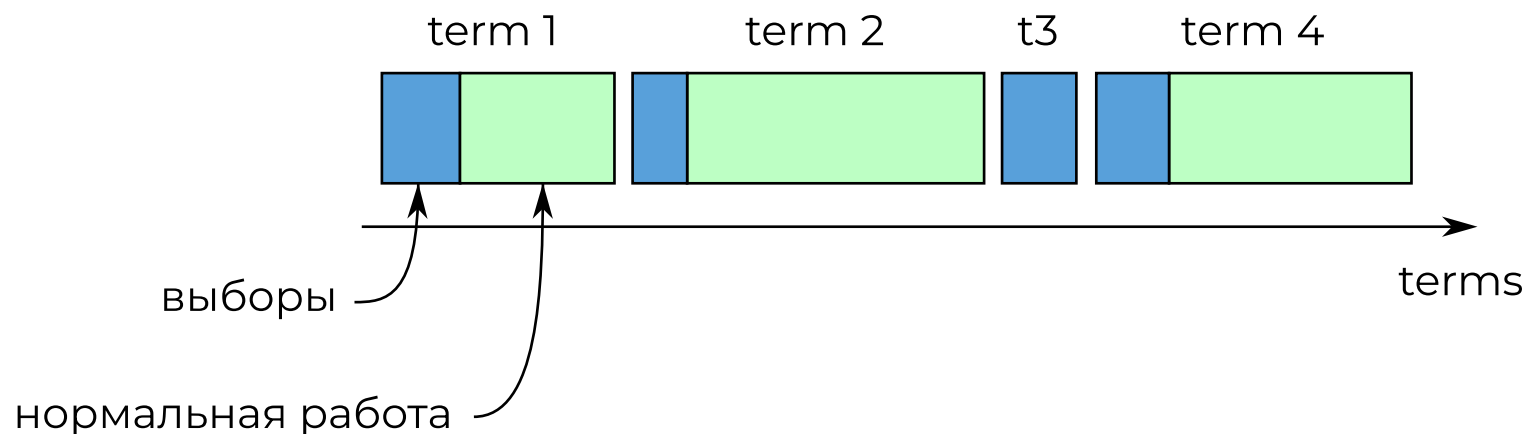
Лидер, фолловер, кандидат



- *Leader* единственный пишет в журнал + пингует окружающих
- *Follower* пассивен, не отправляет никаких запросов
- *Candidate* проводит голосование

Термы

Терм — это отрезок времени неопределенной длины. Он начинается с выборов, после которых единственный лидер управляет кластером.



Выборы лидера

Игра 1



i1, начинайте выборы



Raft State

raft_id
<i>i1</i>

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>													

Raft Log

index:														
term:														
value:														
applied:														

RequestVoteRPC

term:	<i>t1</i>
candidateId:	<i>i1</i>
lastLogIndex:	<i>0</i>
lastLogTerm:	<i>t0</i>

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✓							

i2-i8: «OK».

RequestVoteRPC

term:	<i>t1</i>
candidateId:	<i>i1</i>
lastLogIndex:	<i>0</i>
lastLogTerm:	<i>t0</i>

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✓	✓						

Raft State

raft_id
<i>i2</i>

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>													

Raft Log

index:														
term:														
value:														
applied:														

i1: «Y-xyy!»

Raft State

raft_id
<i>i1</i>

Terms



term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>													

Raft Log

index:														
term:														
value:														
applied:														

RequestVoteRPC

term:	<i>t1</i>
candidateId:	<i>i1</i>
lastLogIndex:	<i>0</i>
lastLogTerm:	<i>t0</i>

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✓	✓	✓	✓	✓	✓	✓	✓

Репликация журнала

Игра 2



іІ, заповняйте raft-журнал

raft_id

i1

Terms

term:

t1

t2

t3

t4

t5

t6

t7

t8

t9

t10

t11

t12

t13

t14

vote:

i1

Raft Log

index:

1

2

3

4

5

6

7

8

term:

t1

t1

t1

...

...

...

...

...

value:

H

i

g

h

L

o

a

d

applied:

i1: «Заперсистьте!»

Raft State

raft_id

i1

Terms



term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>													

Raft Log

index:	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>						
term:	<i>t1</i>	<i>t1</i>	<i>t1</i>						
value:	<i>H</i>	<i>i</i>	<i>g</i>	<i>h</i>	<i>L</i>	<i>o</i>	<i>a</i>	<i>d</i>						
applied:														

AppendEntriesRPC

term:	<i>t1</i>
leaderId:	<i>i1</i>
index:	<i>1</i> <i>2</i> <i>3</i> <i>4</i>
value:	<i>H</i> <i>i</i> <i>g</i> <i>h</i>
leaderCommit:	<i>0</i>

Response

	<i>i1</i>	i2	i3	i4	i5	i6	i7	i8
ok	✓							

i2-i8: «OK».

AppendEntriesRPC

term:	<i>t1</i>
leaderId:	<i>i1</i>
index:	<i>1</i> <i>2</i> <i>3</i> <i>4</i>
value:	<i>H</i> <i>i</i> <i>g</i> <i>h</i>
leaderCommit:	<i>0</i>

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✓	✓						

Raft State

raft_id
<i>i2</i>

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>													

Raft Log


index:	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>										
term:	<i>t1</i>	<i>t1</i>	<i>t1</i>	...										
value:	<i>H</i>	<i>i</i>	<i>g</i>	<i>h</i>										
applied:														

i1: «Отлично!»

Raft State

raft_id
<i>i1</i>

Terms



term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>													

Raft Log

index:	1	2	3	4	5	6	7	8						
term:	<i>t1</i>	<i>t1</i>	<i>t1</i>						
value:	<i>H</i>	<i>i</i>	<i>g</i>	<i>h</i>	<i>L</i>	<i>o</i>	<i>a</i>	<i>d</i>						
applied:	✓	✓	✓	✓										

AppendEntriesRPC

term:

<i>t1</i>

leaderId:

<i>i1</i>

index:

1	2	3	4
---	---	---	---

value:

<i>H</i>	<i>i</i>	<i>g</i>	<i>h</i>
----------	----------	----------	----------

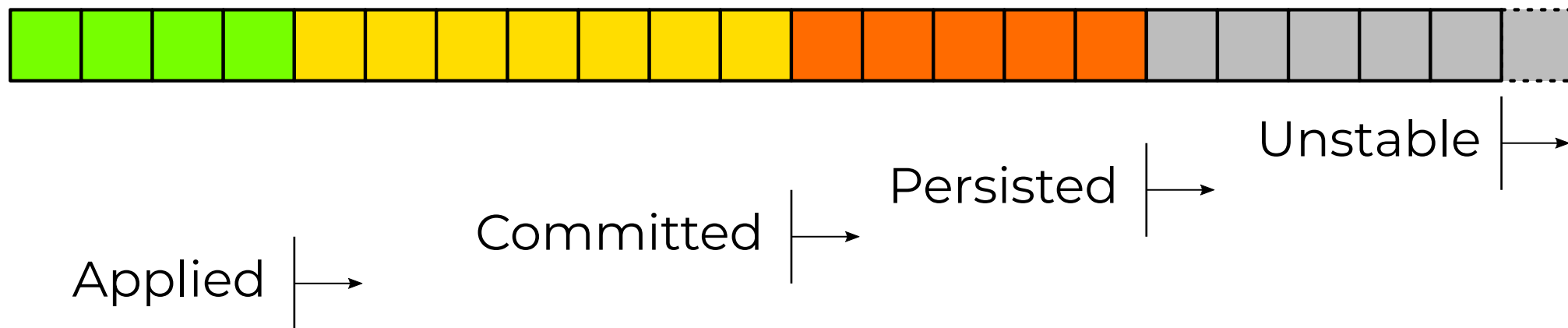
leaderCommit:

0

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✓	✓	✓	✓	✓	✓	✓	✓

Состояние записей



i1: «Реплицируйтесь!»

Raft State

raft_id
<i>i1</i>

Terms



term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>													

Raft Log

index:	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>						
term:	<i>t1</i>	<i>t1</i>	<i>t1</i>						
value:	<i>H</i>	<i>i</i>	<i>g</i>	<i>h</i>	<i>L</i>	<i>o</i>	<i>a</i>	<i>d</i>						
applied:	✓	✓	✓	✓										

AppendEntriesRPC

term:	<i>t1</i>
leaderId:	<i>i1</i>
index:	<i>5</i> <i>6</i> <i>7</i> <i>8</i>
value:	<i>L</i> <i>o</i> <i>a</i> <i>d</i>
leaderCommit:	<i>4</i>

Response

	<i>i1</i>	i2	i3	i4	i5	i6	i7	i8
ok	✓							

i2, i3, сохраняйте записи

AppendEntriesRPC

term:

t1

leaderId:

i1

index:

5 | 6 | 7 | 8

value:

L | o | a | d

leaderCommit:

4

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✓	✓	✓					

Raft State

raft_id

i2

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>													

Raft Log

index:	1	2	3	4	5	6	7	8						
term:	<i>t1</i>	t1	t1							
value:	H	i	g	h	L	o	a	d						
applied:	✓	✓	✓	✓										

i4, "потеряйте" сообщение

AppendEntriesRPC

term:

t1

leaderId:

i1

index:

5 | 6 | 7 | 8

value:

L | o | a | d

leaderCommit:

4

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✓	✓	✓					

Raft State

raft_id

i4

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	i1													

Raft Log

index:	1	2	3	4										
term:	t1	t1	t1	...										
value:	H	i	g	h										
applied:														

Терм без лидера

Игра 3



i4 и i7, вы "оффлайн"

Переверните ваши листки

i8, начинайте выборы

Raft State

raft_id
<i>i8</i>

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>	<i>i8</i>												

Raft Log

index:	1	2	3	4										
term:	<i>t1</i>	<i>t1</i>	<i>t1</i>	...										
value:	<i>H</i>	<i>i</i>	<i>g</i>	<i>h</i>										
applied:														

RequestVoteRPC

term:	<i>t2</i>
candidateId:	<i>i8</i>
lastLogIndex:	<i>4</i>
lastLogTerm:	<i>t1</i>

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok								✓

i2, i3: «He-a».

RequestVoteRPC

term:	<i>t2</i>
candidateId:	<i>i8</i>
lastLogIndex:	<i>4</i>
lastLogTerm:	<i>t1</i>

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok		X						✓

S may only vote for L if:

$L.lastLogTerm > S.lastLogTerm$ or

$(L.lastLogTerm == S.lastLogTerm$ and

$L.lastLogIndex \geq S.lastLogIndex)$

Raft State

raft_id
<i>i2</i>

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>	-												

Raft Log

index:	1	2	3	4	5	6	7	8						
term:	<i>t1</i>	<i>t1</i>	<i>t1</i>						
value:	<i>H</i>	<i>i</i>	<i>g</i>	<i>h</i>	<i>L</i>	<i>o</i>	<i>a</i>	<i>d</i>						
applied:	✓	✓	✓	✓										

i8: «😓».

Raft State

raft_id
<i>i8</i>

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>	<i>i8</i>												

Raft Log

index:	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>										
term:	<i>t1</i>	<i>t1</i>	<i>t1</i>	...										
value:	<i>H</i>	<i>i</i>	<i>g</i>	<i>h</i>										
applied:														

RequestVoteRPC

term:	<i>t2</i>
candidateId:	<i>i8</i>
lastLogIndex:	<i>4</i>
lastLogTerm:	<i>t1</i>

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok		X	X		✓	✓	✓	✓

Перезапись журнала

Игра 4



іІ все еще "оффлайн"

Потерпите, через 4 слайда вернетесь.

i4, начинайте выборы

Raft State

raft_id
<i>i4</i>

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>	-	<i>i4</i>											

Raft Log

index:	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>										
term:	<i>t1</i>	<i>t1</i>	<i>t1</i>	...										
value:	<i>H</i>	<i>i</i>	<i>g</i>	<i>h</i>										
applied:														

RequestVoteRPC

term:	<i>t3</i>
candidateId:	<i>i4</i>
lastLogIndex:	<i>4</i>
lastLogTerm:	<i>t1</i>

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok				✓				

Вжух, и i4 теперь лидер

Raft State

raft_id

i4

Terms

term:

t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	<i>i1</i>	–	<i>i4</i>										

Raft Log

index:

1	2	3	4										
term:	<i>t1</i>	<i>t1</i>	<i>t1</i>	...									
value:	<i>H</i>	<i>i</i>	<i>g</i>	<i>h</i>									
applied:													

RequestVoteRPC

term:

<i>t3</i>	
candidateId:	<i>i4</i>
lastLogIndex:	4
lastLogTerm:	<i>t1</i>

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok		X	X	✓	✓	✓	✓	✓

[illegible]

i5-i8, обработайте запрос

AppendEntriesRPC

term:

t3

leaderId:

i4

index:

5 | 6 | 7 | 8

value:

| v | o | l

leaderCommit:

0

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok				✓	✓			

Raft State

raft_id

i5

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	i1	–	i4											

Raft Log

index:	1	2	3	4	5	6	7	8						
term:	t1	t1	t1	...	t3	t3	t3	...						
value:	H	i	g	h	␣	v	o	l						
applied:	✓	✓	✓	✓										

i1, возвращайтесь онлайн

AppendEntriesRPC

term:	t_3
leaderId:	i_4
index:	5 6 7 8
value:	v o l
leaderCommit:	0

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✓			✓	✓	✓	✓	✓

Raft State														
														raft_id
														i1
Terms														
term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	i1	-	-											
Raft Log														
index:	1	2	3	4	5	6	7	8	5	6	7	8		
term:	t1	t1	t1	t3	t3	t3	...		
value:	H	i	g	h	L	o	a	d	_	v	o	l		
applied:	✓	✓	✓	✓										

i4 получает ответ



Raft State

raft_id

i4



Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	i1	-	i4											

Raft Log

index:	1	2	3	4	5	6	7	8	9	10	11	12		
term:	t1	t1	t1	...	t3	t3	t3		
value:	H	i	g	h	␣	v	o	l	t	a	g	e		
applied:	✓	✓	✓	✓	✓	✓	✓	✓						

AppendEntriesRPC

term:	t3
leaderId:	i4
index:	5 6 7 8
value:	 v o l
leaderCommit:	0

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✓	✓	✓	✓	✓	✓	✓	✓

Факультатив

- Динамическое изменение топологии.
- Pre-vote.
- Applied индекс может обгонять persisted.

Материалы

Слайды:

- Online: <https://rosik.github.io/2023-highload>
- PDF: [slides.pdf](#)
- Раздатка: [form.pdf](#)

Picodata: <https://picodata.io/>, [@picodataru](#)

Raft: <https://raft.github.io/>

Обратная связь:



<https://conf.ontico.ru/online/shl2023/details/4937163>

