Как при помощи бумаги, карандаша и алгоритма Raft достичь консенсуса

Ярослав Дынников

Picodata

HL++ 2023

HighLoad++

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

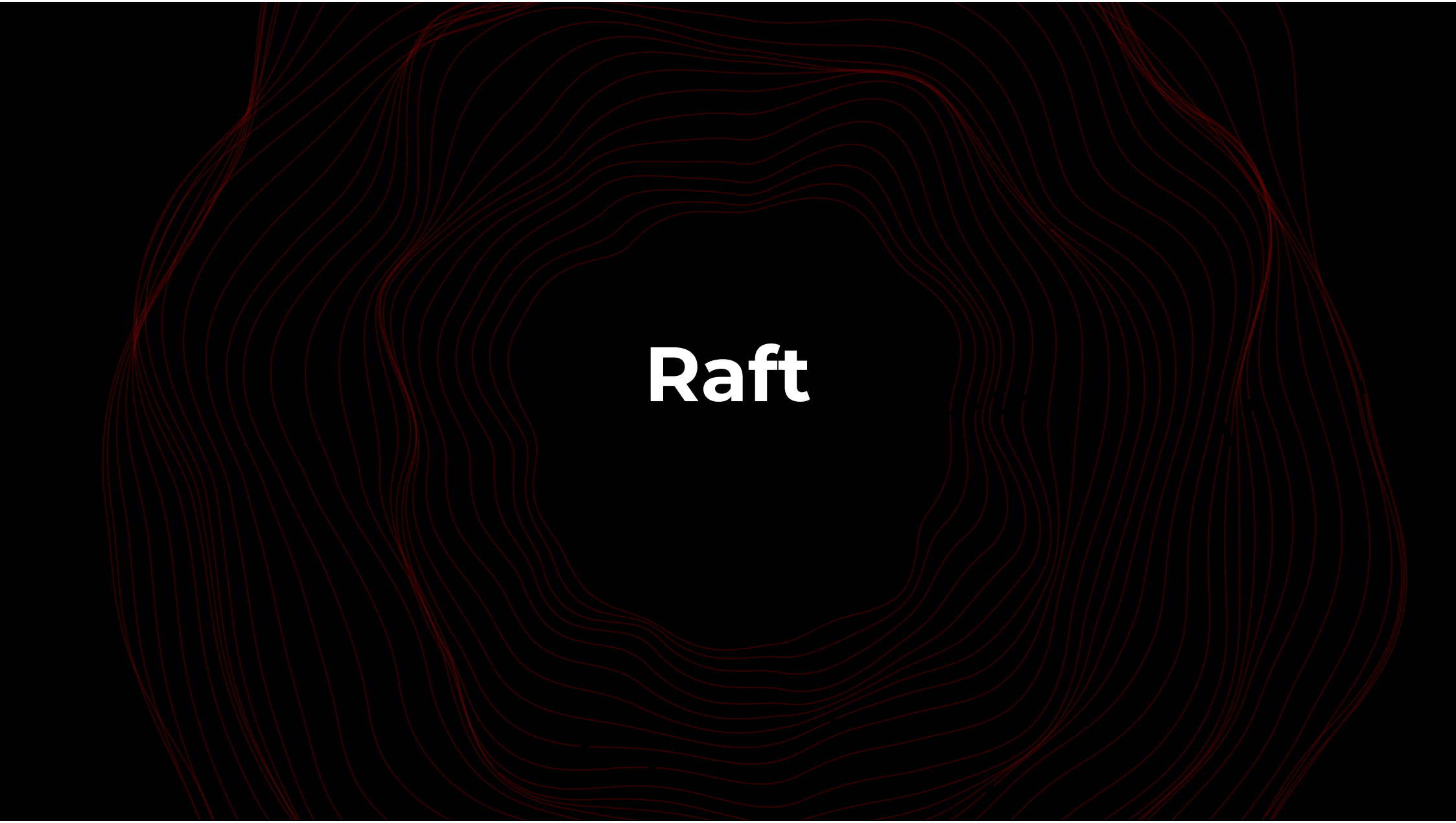
Кластер — это группа процессов, работающих совместно и представляющихся пользователю единым компьютерным ресурсом.

Задача

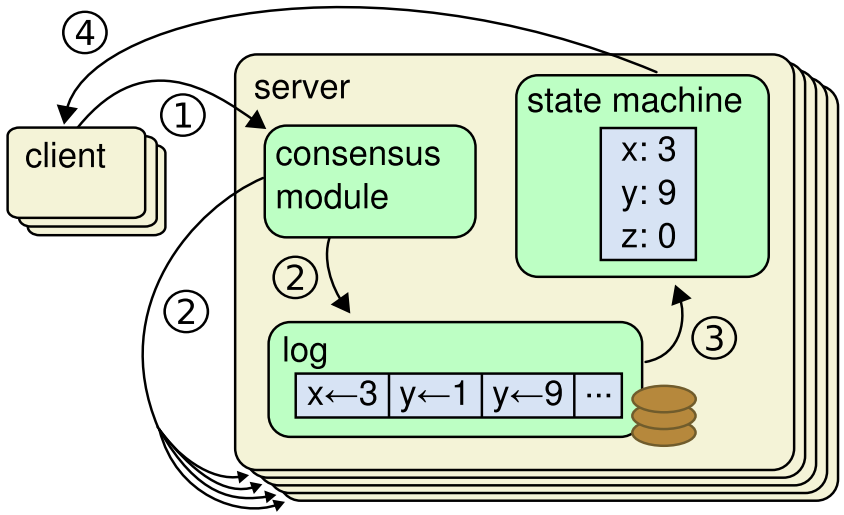
- Есть несколько серверов.
- Надо достичь консенсуса.
- В ненадежной сети.

Решение — Raft

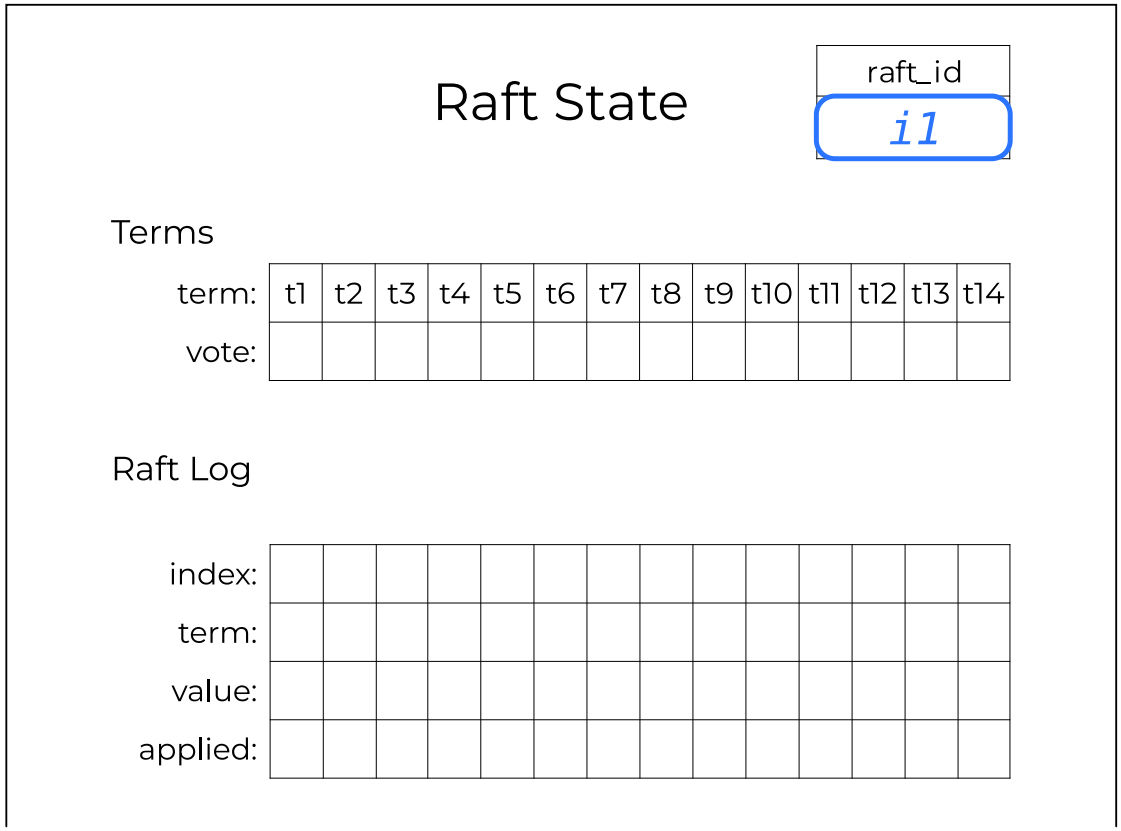
- In search of Understandable Consensus Algorithm.
- Diego Ongaro and John Ousterhout. Stanford University.
- https://raft.github.io


Raft

Реплицируемый конечный автомат

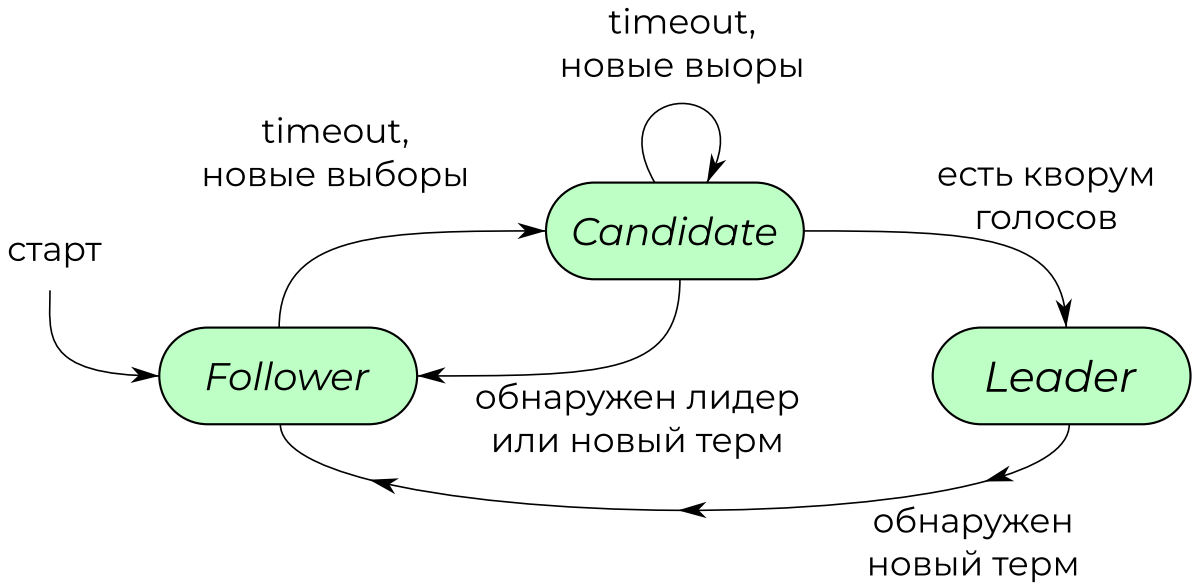


Персистентное хранилище



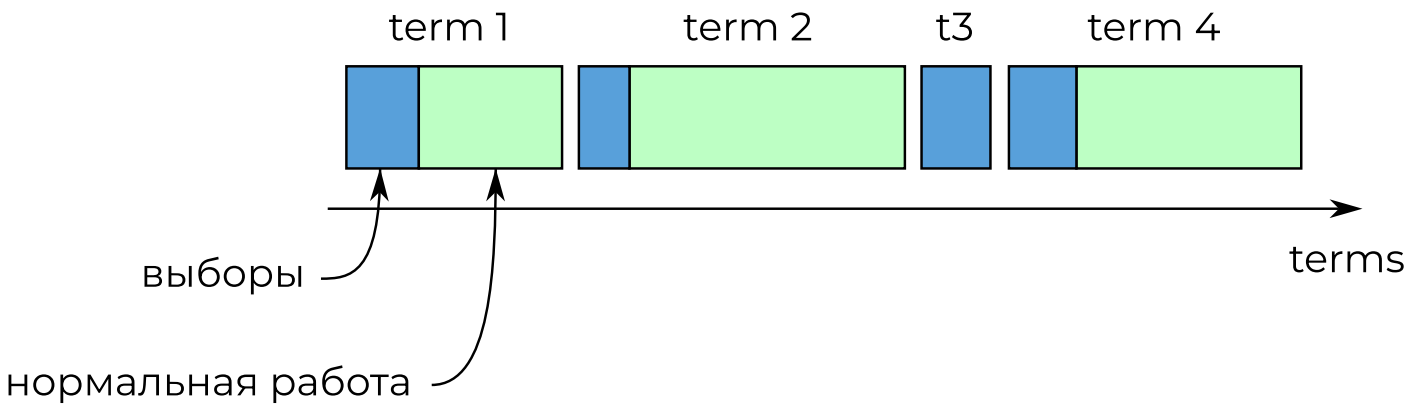
Лидер, фолловер, кандидат

- *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
i1

Terms

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

Raft Log

index:														
term:														
value:														
applied:														

RequestVoteRPC

term:	*t1*
candidateId:	*i1*
lastLogIndex:	*0*
lastLogTerm:	*t0*

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✔							

i2-i8: «Ок»

RequestVoteRPC

term:	*t1*
candidateId:	*i1*
lastLogIndex:	*0*
lastLogTerm:	*t0*

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:	*i1*													

Raft Log

index:														
term:														
value:														
applied:														

i1: «У-хуу!»

Raft State

raft_id
i1

Terms

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

Raft Log

index:														
term:														
value:														
applied:														

RequestVoteRPC

term:	*t1*
candidateId:	*i1*
lastLogIndex:	*0*
lastLogTerm:	*t0*

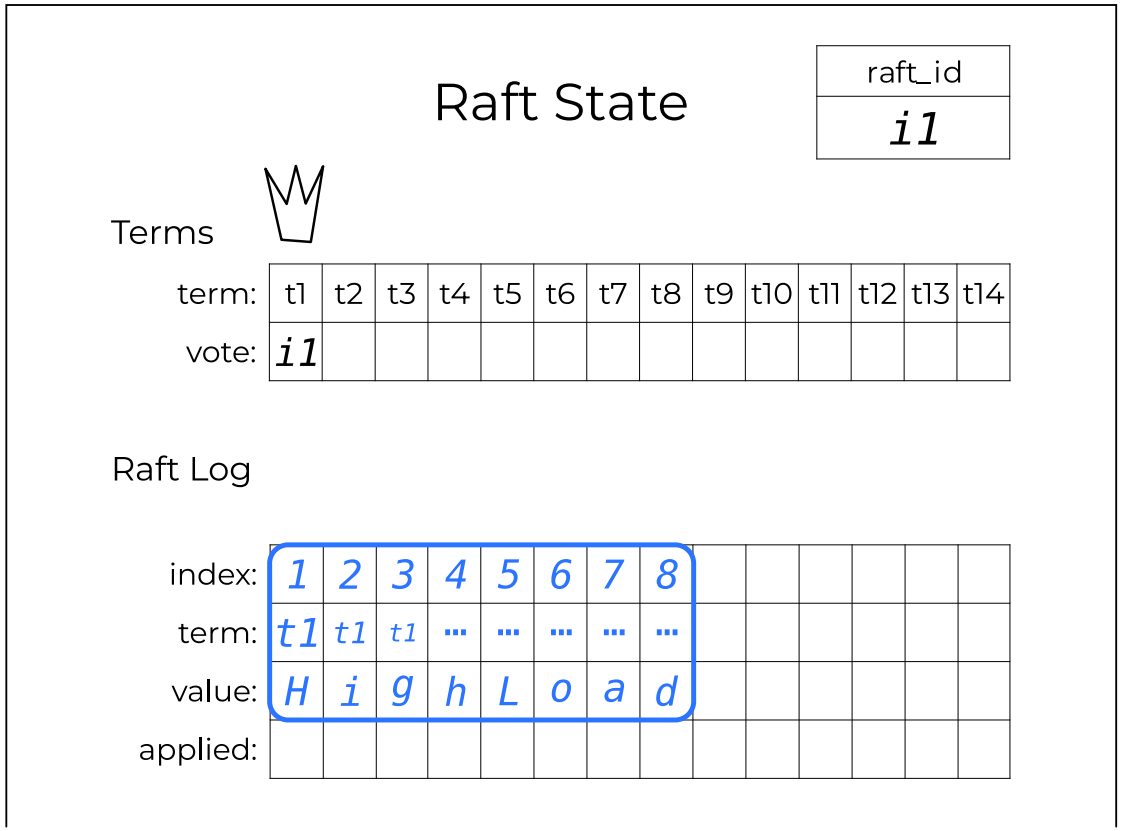
Response

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

Репликация журнала

Игра 2

i1, заполняйте raft-журнал



i1: «Заперсистьте!»

Raft State

raft_id
i1

Terms

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

AppendEntriesRPC

term:	*t1*
leaderId:	*i1*
index:	*1* | *2* | *3* | *4*
value:	*H* | *i* | *g* | *h*
leaderCommit:	*0*

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok	✔							

i2-i8: «Ок»

AppendEntriesRPC

term:	*t1*
leaderId:	*i1*
index:	*1 | 2 | 3 | 4*
value:	*H | i | g | h*
leaderCommit:	*0*

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:	*i1*													

Raft Log

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

i1: «Отлично!»

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*						
term:	*t1*	*t1*	*t1*	…	…	…	…	…						
value:	*H*	*i*	*g*	*h*	*L*	*o*	*a*	*d*						
applied:	✔	✔	✔	✔										

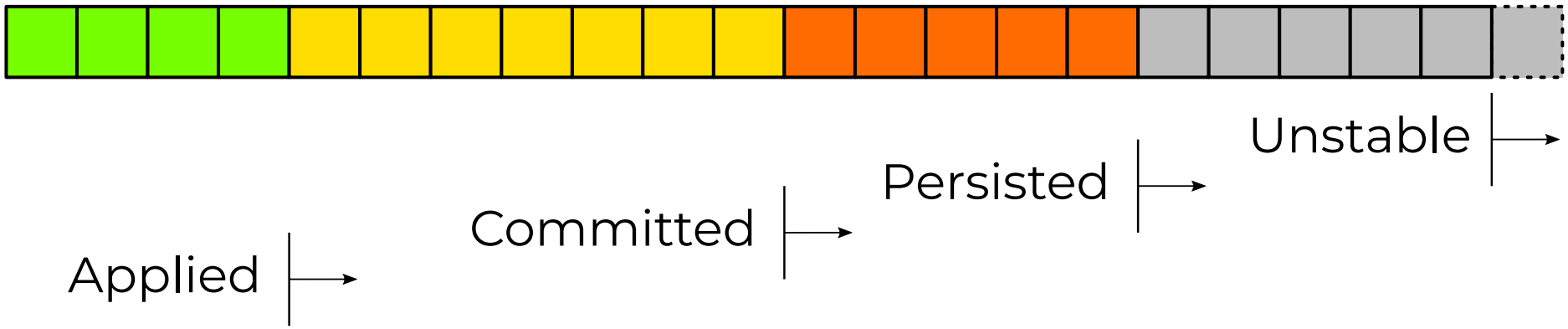
AppendEntriesRPC

term:	*t1*
leaderId:	*i1*
index:	*1* | *2* | *3* | *4*
value:	*H* | *i* | *g* | *h*
leaderCommit:	*0*

Response

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

Состояние записей



i1: «Реплицируйтесь!»

Raft State

raft_id
i1

Terms

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

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	✔							

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:	*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:	✔	✔	✔	✔										

HighLoad++

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 i1, вы offline

Переверните ваши листки

i8, начинайте выборы

Raft State

raft_id
i8

Terms

term:	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
vote:	*i1*	*i8*												

Raft Log

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

RequestVoteRPC

term:	*t2*
candidateId:	*i8*
lastLogIndex:	*4*
lastLogTerm:	*t1*

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok								✔

i2, i3: «He-a»

RequestVoteRPC

term:	*t2*
candidateId:	*i8*
lastLogIndex:	*4*
lastLogTerm:	*t1*

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok		✗						✔

```
S may only vote for L if:
L.lastLogTerm > S.lastLogTerm or
  (L.lastLogTerm == S.lastLogTerm and
   L.lastLogIndex ≥ S.lastLogIndex)
```

Raft State

raft_id
i2

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:	✔	✔	✔	✔									

i8: «😢»

Raft State

raft_id
i8

Terms

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11	t12	t13	t14
term:														
vote:	*i1*	*i8*												

Raft Log

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

RequestVoteRPC

term:	*t2*
candidateId:	*i8*
lastLogIndex:	*4*
lastLogTerm:	*t1*

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok		✗	✗		✔	✔	✔	✔

Перезапись журнала

Игра 4

i1 все еще offline

Потерпите, через 4 слайда вернетесь.

i4, начинайте выборы

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*									
term:	*t1*	*t1*	*t1*	⋯									
value:	*H*	*i*	*g*	*h*									
applied:													

RequestVoteRPC

term:	*t3*
candidateId:	*i4*
lastLogIndex:	*4*
lastLogTerm:	*t1*

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok				✔				

Вжух, и i4 — лидер

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*									
term:	*t1*	*t1*	*t1*	⋯									
value:	*H*	*i*	*g*	*h*									
applied:													

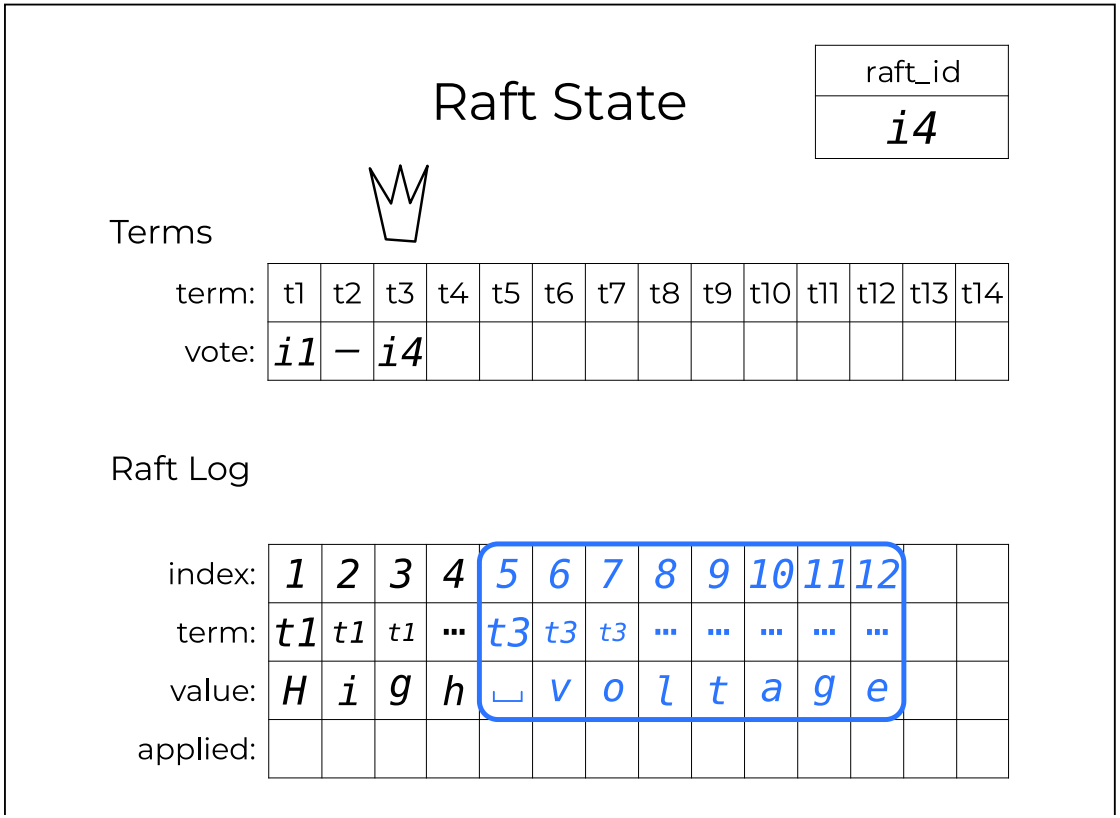
RequestVoteRPC

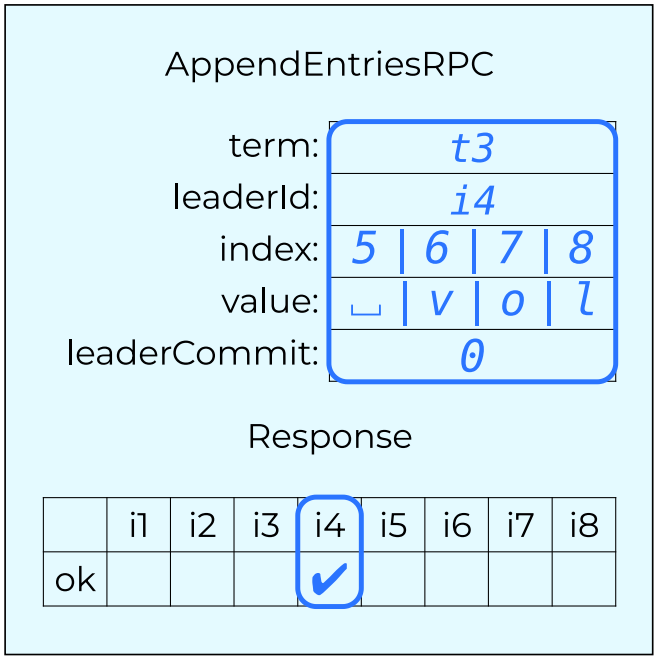
term:	*t3*
candidateId:	*i4*
lastLogIndex:	*4*
lastLogTerm:	*t1*

Response

	i1	i2	i3	i4	i5	i6	i7	i8
ok		✗	✗	✔	✔	✔	✔	✔

i4, заполняйте raft-журнал





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:	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
i1

Terms

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

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

Raft Log

	1	2	3	4	5	6	7	8	9	10	11	12		
index:														
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
- Снапшоты
- Quorum read

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/hl2023/lectures/5071966

HighLoad++