Решение практических проблем Raft в Tarantool

Что, как и почему

Сергей Петренко
Tarantool

Борис Степаненко
Tarantool

Tarantool

Hydra

Привет!

Меня зовут Сергей Петренко

Привет!

Меня зовут Сергей Петренко

Последние 4 года – программист в Tarantool

Привет!

Меня зовут Сергей Петренко

Последние 4 года – программист в Tarantool

Занимаюсь репликацией

Привет!

Меня зовут Борис

Привет!

Меня зовут Борис

Разрабатываю Tarantool последние полгода

Привет!

Меня зовут Борис

Разрабатываю Tarantool последние полгода

До этого разрабатывал системы информационной безопасности

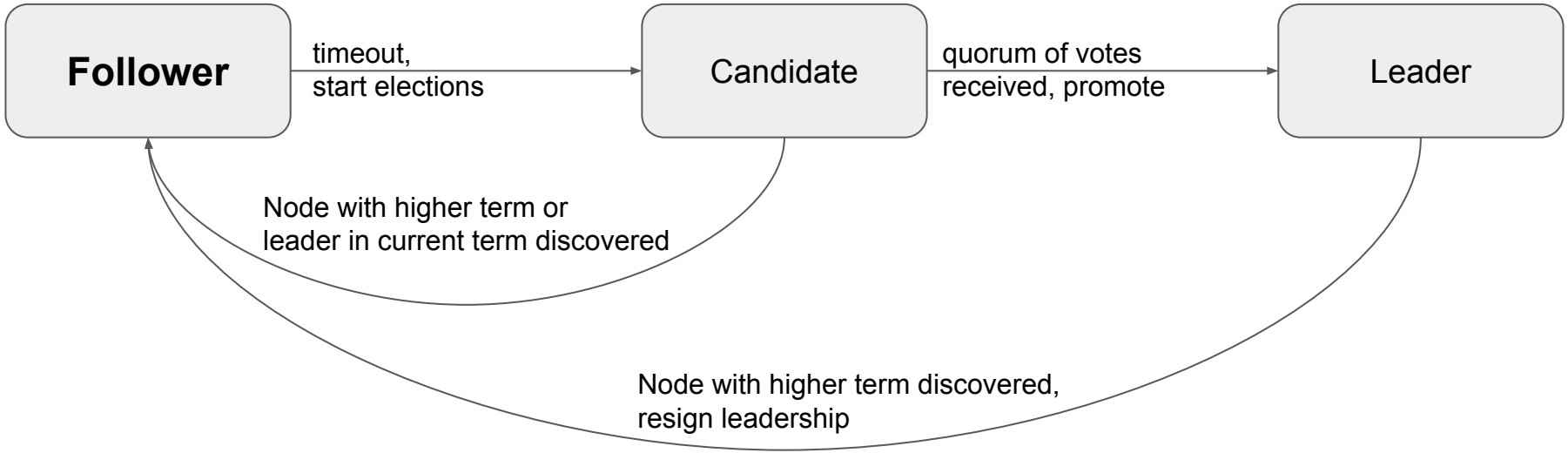
У нас был план и мы его придерживались

- **Raft overview**
  - Journal, Term, LSN
  - Выборы
  - Гарантии
- Raft / Tarantool: особенности
- Надстройки Raft
  - Pre-Vote
  - Split-Vote detection
  - Fencing
  - Manual promotion & Split-Brain detection

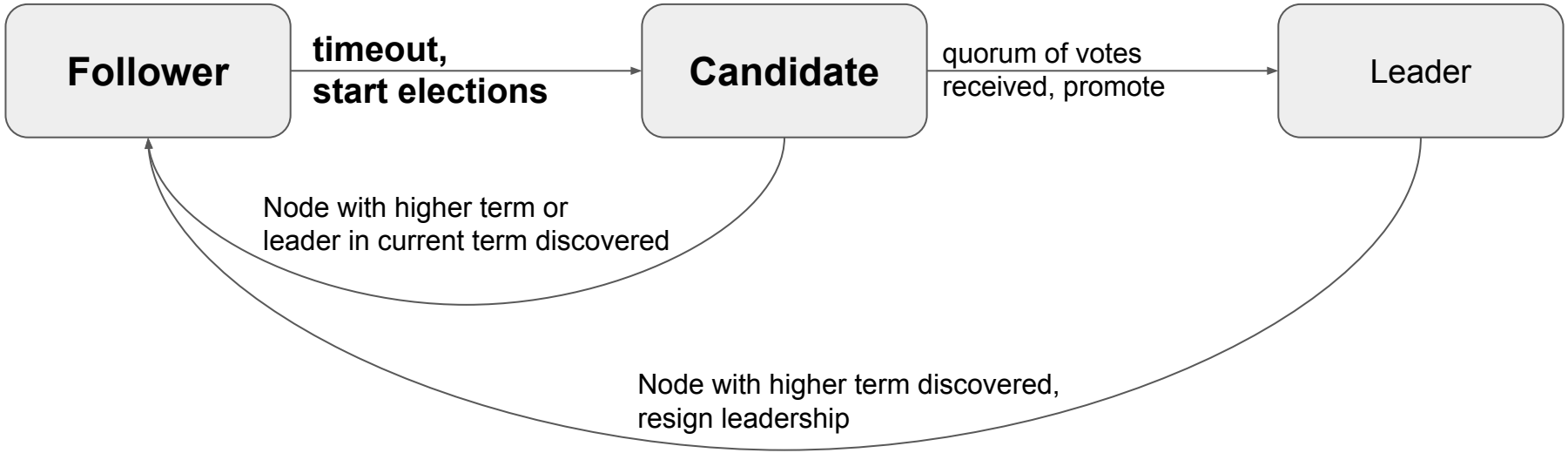
Raft overview

- **Raft** - алгоритм достижения консенсуса в распределённых системах
- Raft достигает консенсуса за счет **выбора единого лидера**, который может изменять состояние системы.
- Лидер выбирается в рамках **term** - пронумерованный отрезок времени
- Лидер может добавлять записи в **журнал (log)**
- Каждая запись в журнале пронумерована **log index/log sequence number (lsn)**

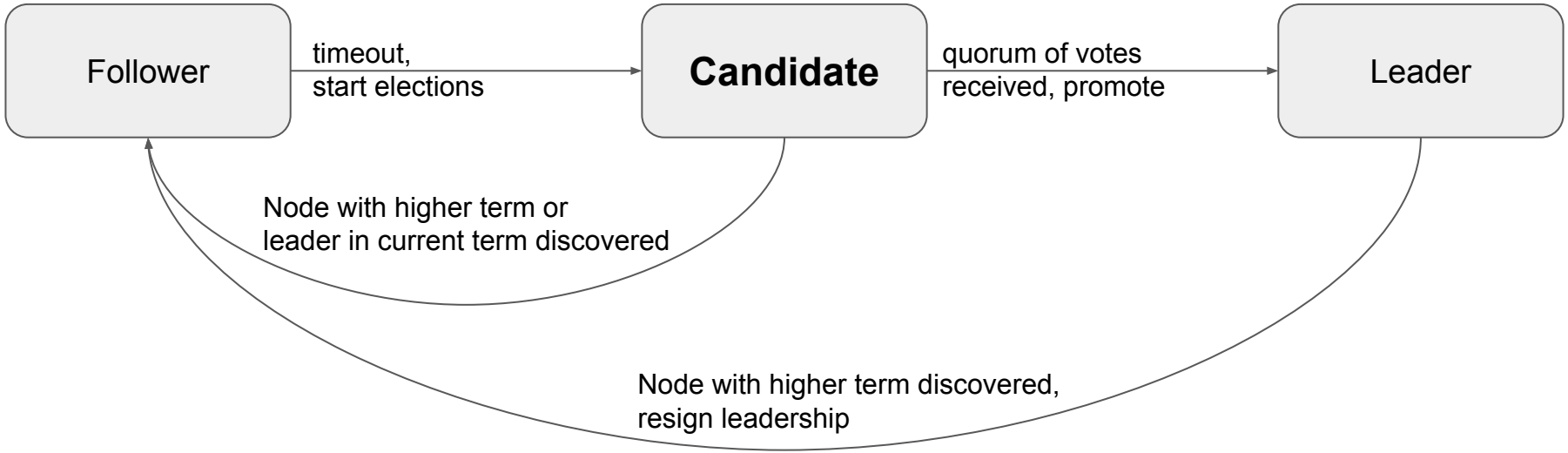
Выборы в RAFT



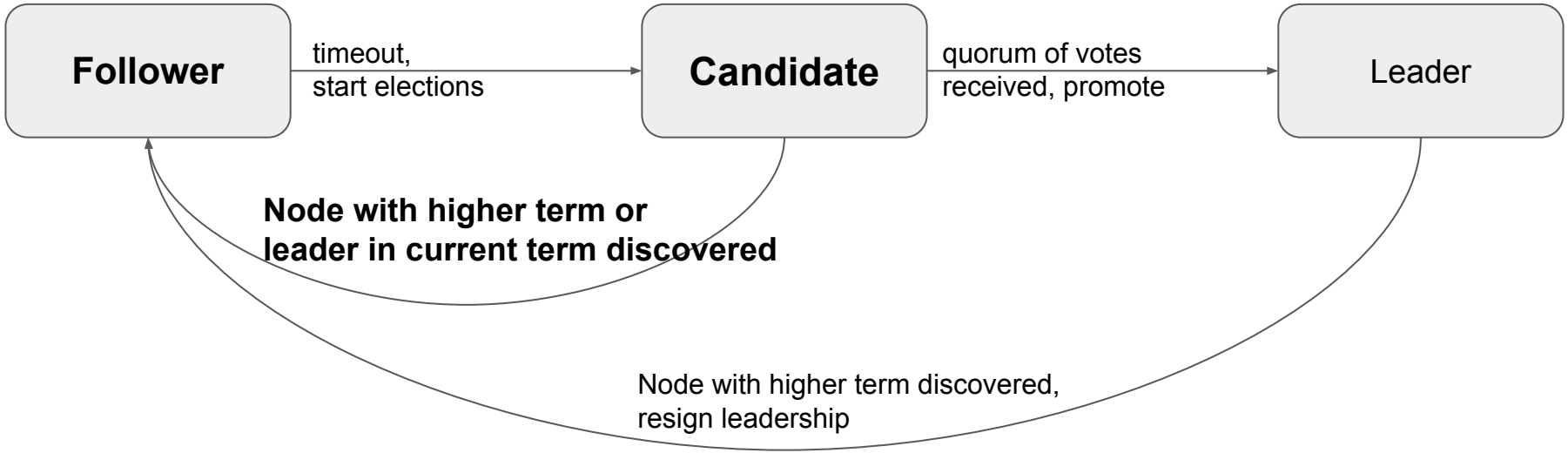
Выборы в RAFT



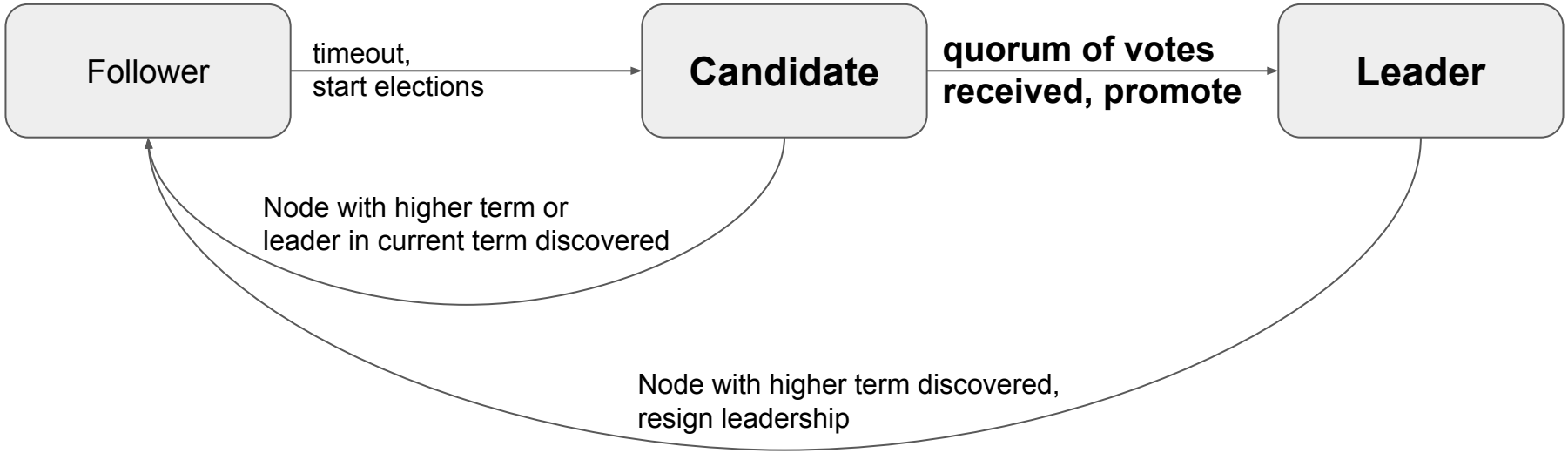
Выборы в RAFT



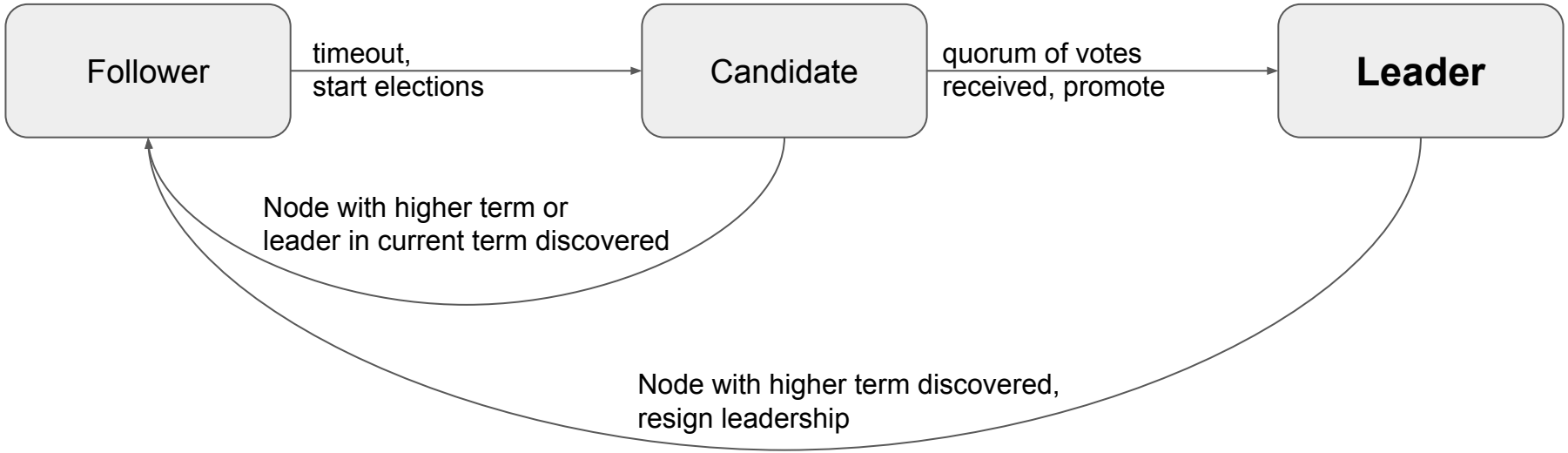
Выборы в RAFT



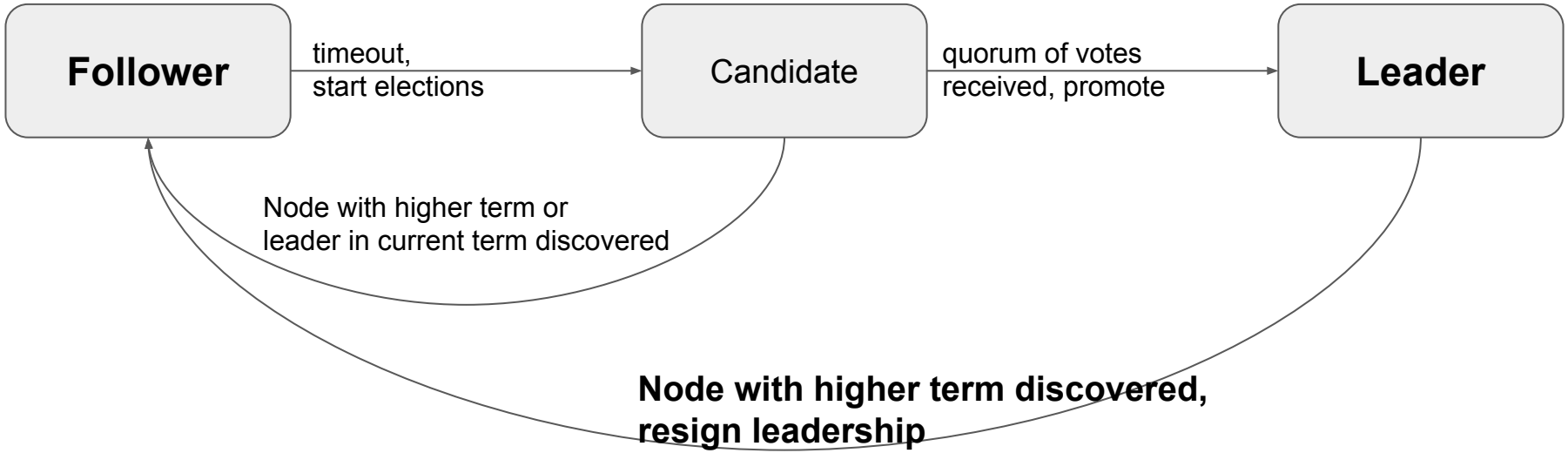
Выборы в RAFT



Выборы в RAFT



Выборы в RAFT



Гарантии

- **Election Safety**: at most one leader can be elected in a given term.

Гарантии

- Election Safety: at most one leader can be elected in a given term.
- **Leader Append-Only**: a leader never overwrites or deletes entries in its log; it only appends new entries.

Гарантии

- Election Safety: at most one leader can be elected in a given term.
- Leader Append-Only: a leader never overwrites or deletes entries in its log; it only appends new entries.
- **Log Matching**: if two logs contain an entry with the same index and term, then the logs are identical in all entries up through the given index.

Гарантии

- Election Safety: at most one leader can be elected in a given term.
- Leader Append-Only: a leader never overwrites or deletes entries in its log; it only appends new entries.
- Log Matching: if two logs contain an entry with the same index and term, then the logs are identical in all entries up through the given index.
- **Leader Completeness**: if a log entry is committed in a given term, then that entry will be present in the logs of the leaders for all higher-numbered terms.

Гарантии

- Election Safety: at most one leader can be elected in a given term.
- Leader Append-Only: a leader never overwrites or deletes entries in its log; it only appends new entries.
- Log Matching: if two logs contain an entry with the same index and term, then the logs are identical in all entries up through the given index.
- Leader Completeness: if a log entry is committed in a given term, then that entry will be present in the logs of the leaders for all higher-numbered terms.
- **State Machine Safety**: if a server has applied a log entry at a given index to its state machine, no other server will ever apply a different log entry for the same index.

Гарантии

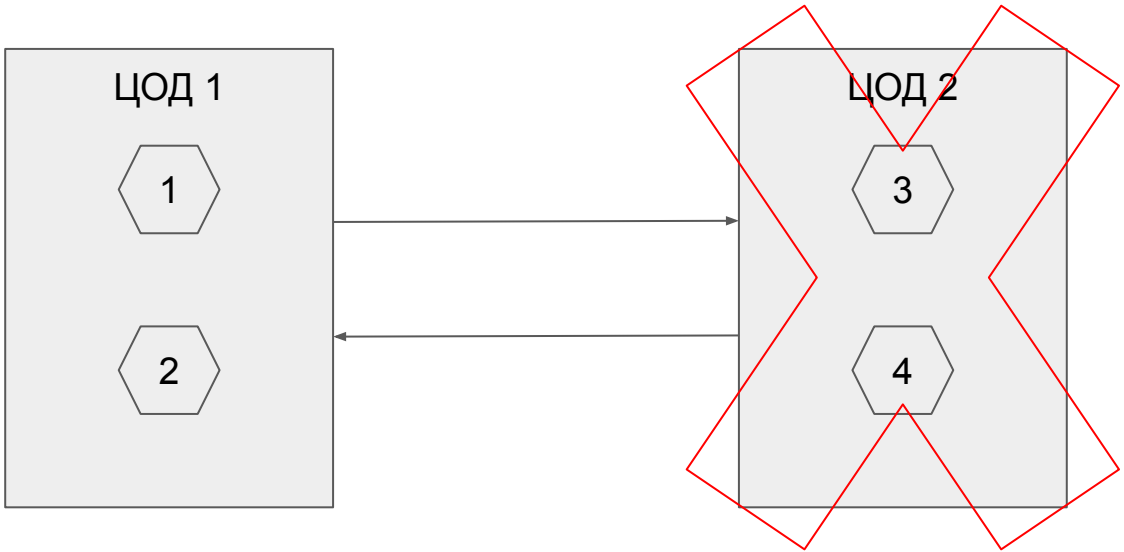
- Election Safety: at most one leader can be elected in a given term.
- Leader Append-Only: a leader never overwrites or deletes entries in its log; it only appends new entries.
- Log Matching: if two logs contain an entry with the same index and term, then the logs are identical in all entries up through the given index.
- Leader Completeness: if a log entry is committed in a given term, then that entry will be present in the logs of the leaders for all higher-numbered terms.
- State Machine Safety: if a server has applied a log entry at a given index to its state machine, no other server will ever apply a different log entry for the same index.

Гарантии

- **Рано или поздно** лидер найдётся (скорее всего)
- **В рамках одного term** есть не более одного лидера

Гарантии

- **Рано или поздно** лидер найдётся (скорее всего)
- **В рамках одного term** есть не более одного лидера



Гарантии

- **Рано или поздно** лидер найдётся (скорее всего)
- **В рамках одного term** есть не более одного лидера

При ограничениях

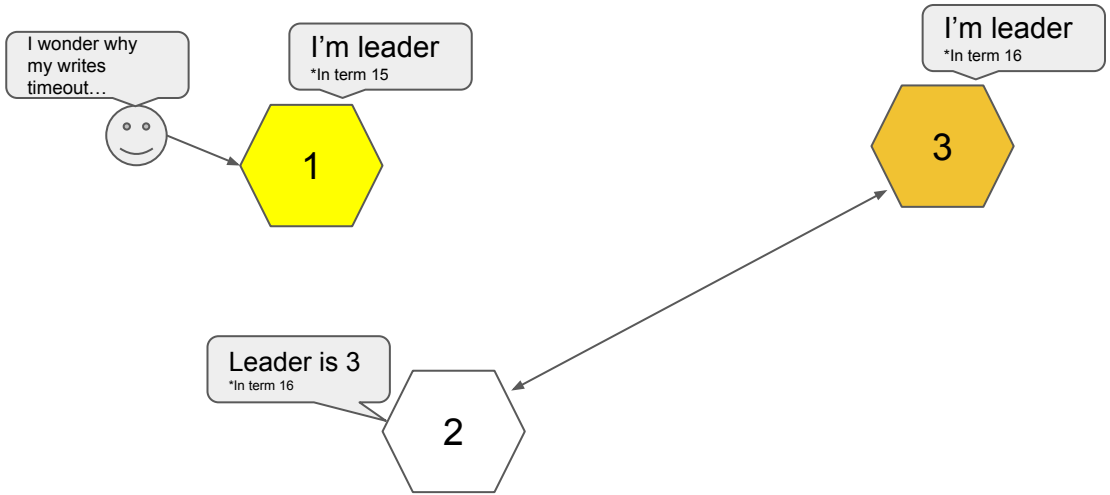
- В кластере есть нода, имеющая необходимое количество соединений, чтобы выиграть выборы
- $T_{ping} << T_{election} << T_{accident}$

Ожидание != реальность

- Рано или поздно != быстро
- Term != время

Ожидание != реальность

- Рано или поздно != быстро
- Term != время



Хочется

- Найти лидера **быстро**
- Не более одного лидера в кластере в **конкретный момент времени**
- Возвращение/подключение новых нод в кластер не вызывает его простоя

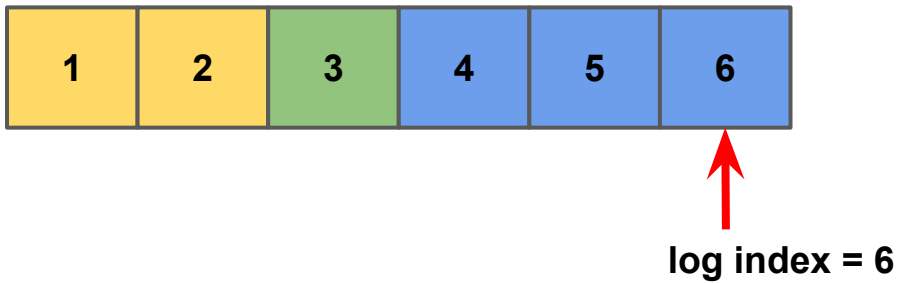
Почему хочется?

On November 2, 2020, Cloudflare had an incident that impacted the availability of the API and dashboard for six hours and 33 minutes. During this incident, the

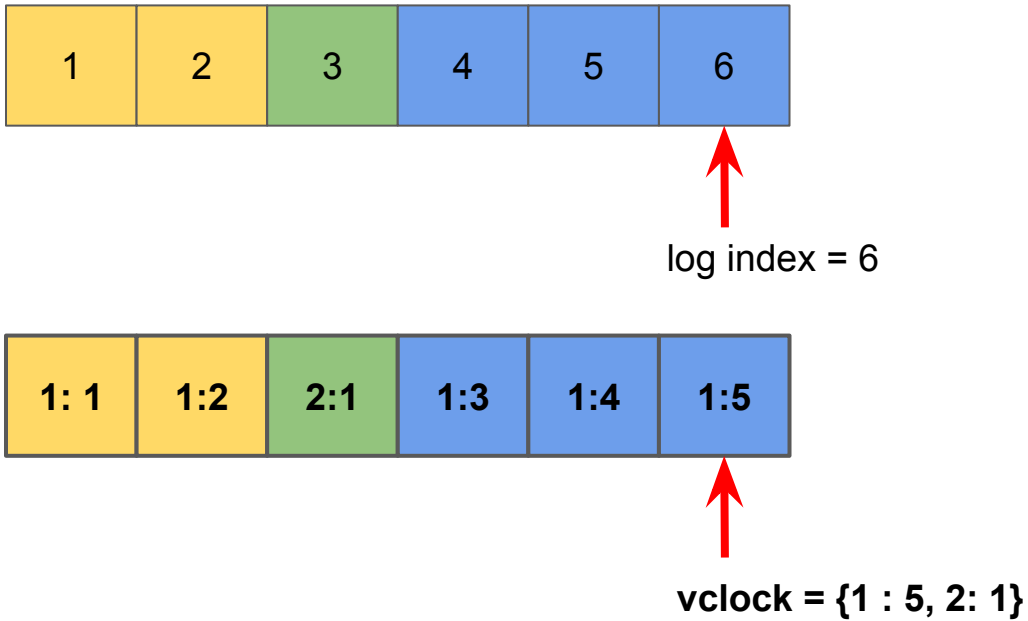
У нас был план и мы его придерживались

- ✓ Raft overview
  - Journal, Term, LSN
  - Выборы
  - Гарантии
- **Raft / Tarantool: особенности**
- Надстройки Raft
  - Pre-Vote
  - Split-Vote detection
  - Fencing
  - Manual promotion & Split-Brain detection

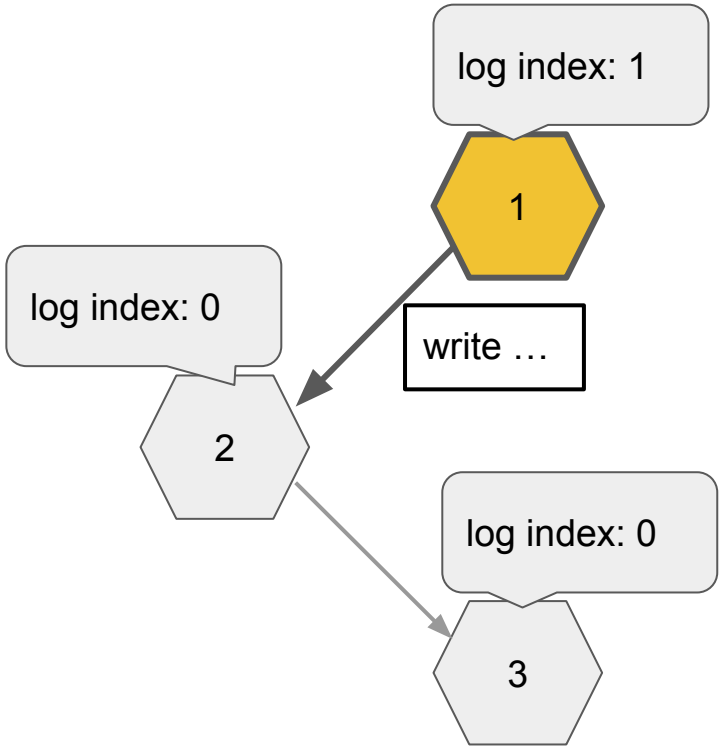
Мультиплексированный журнал

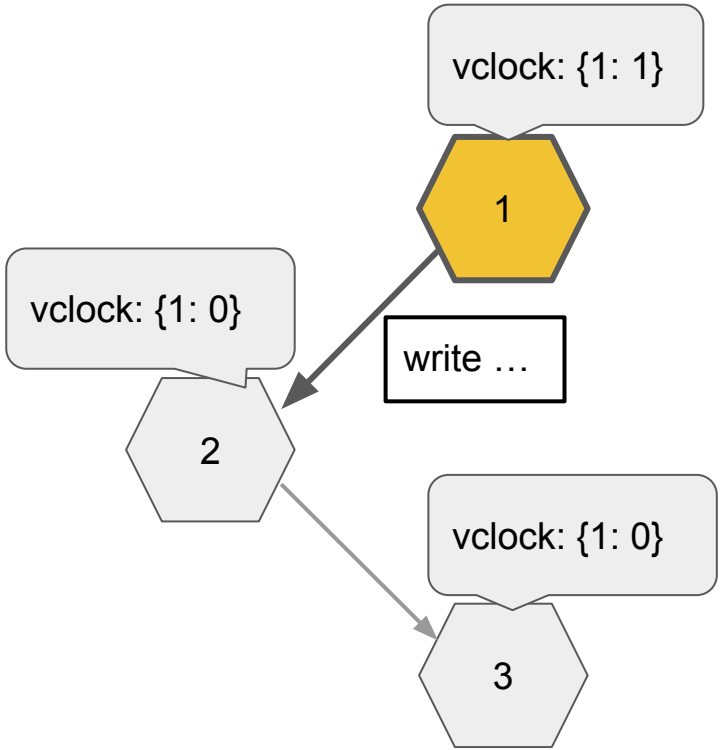


Мультиплексированный журнал

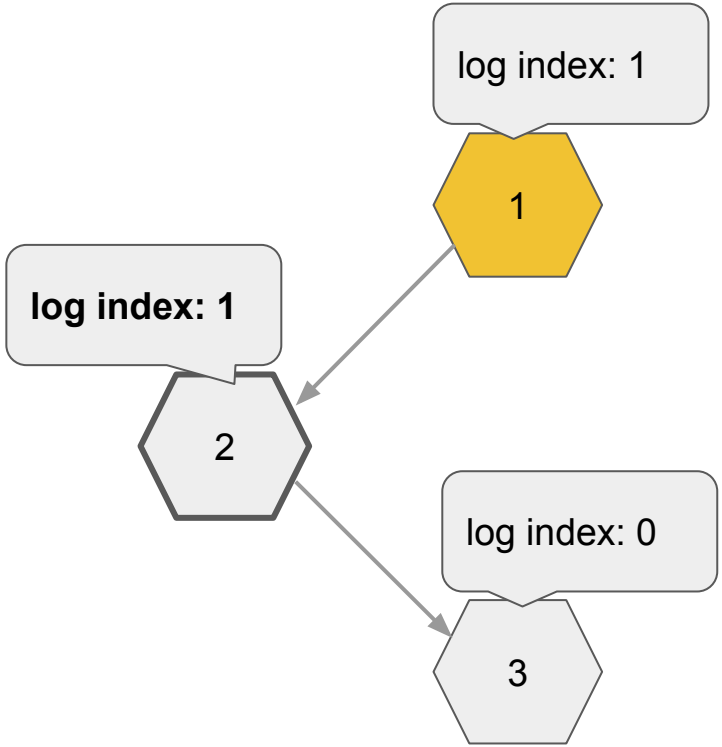


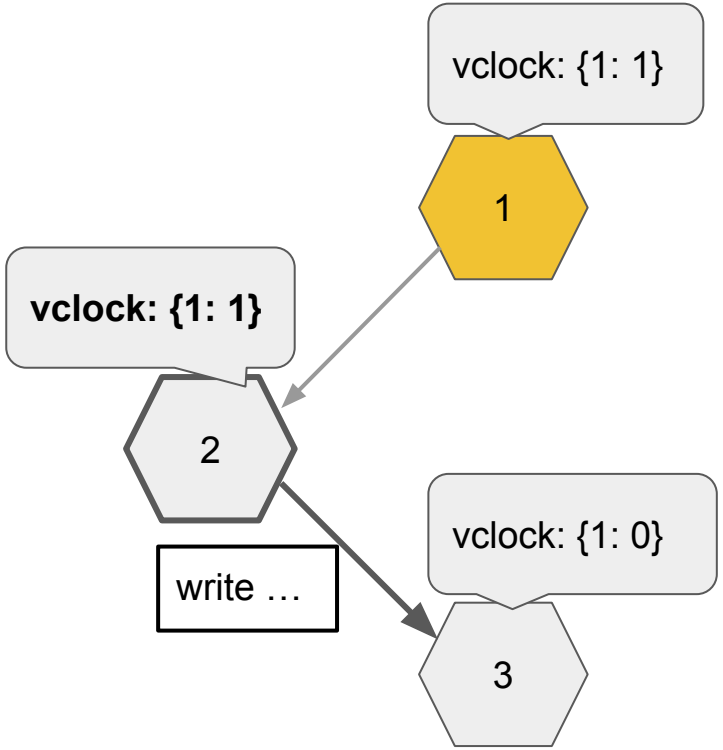
Рассылка данных от всех ко всем



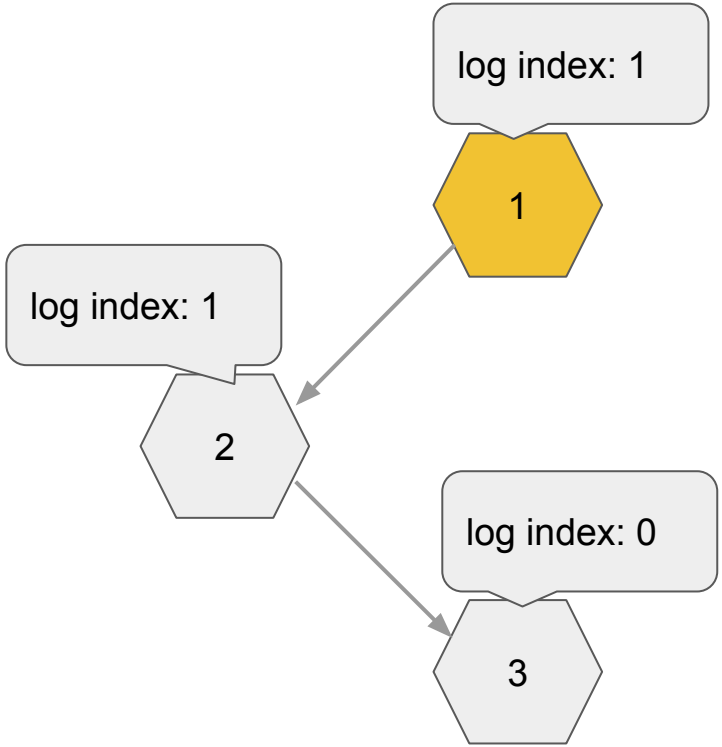


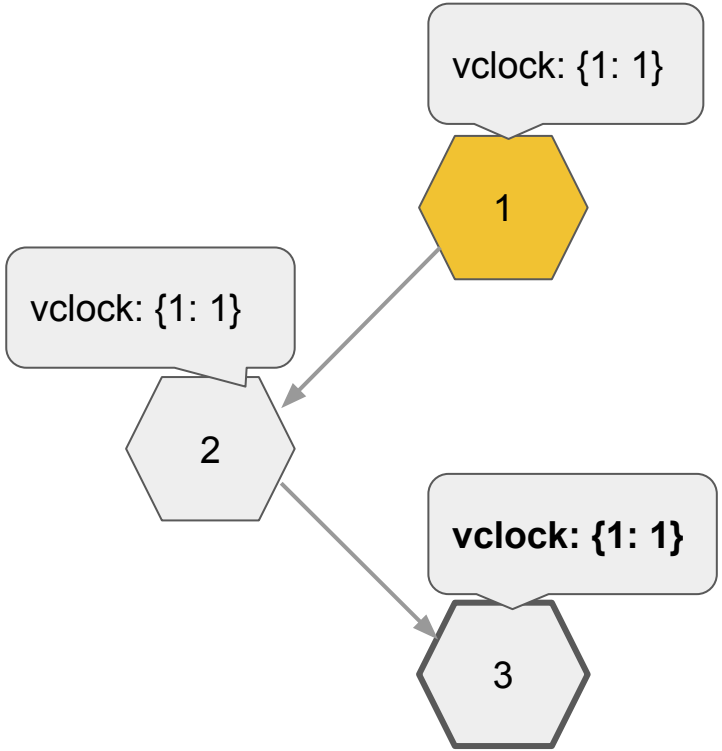
Рассылка данных от всех ко всем



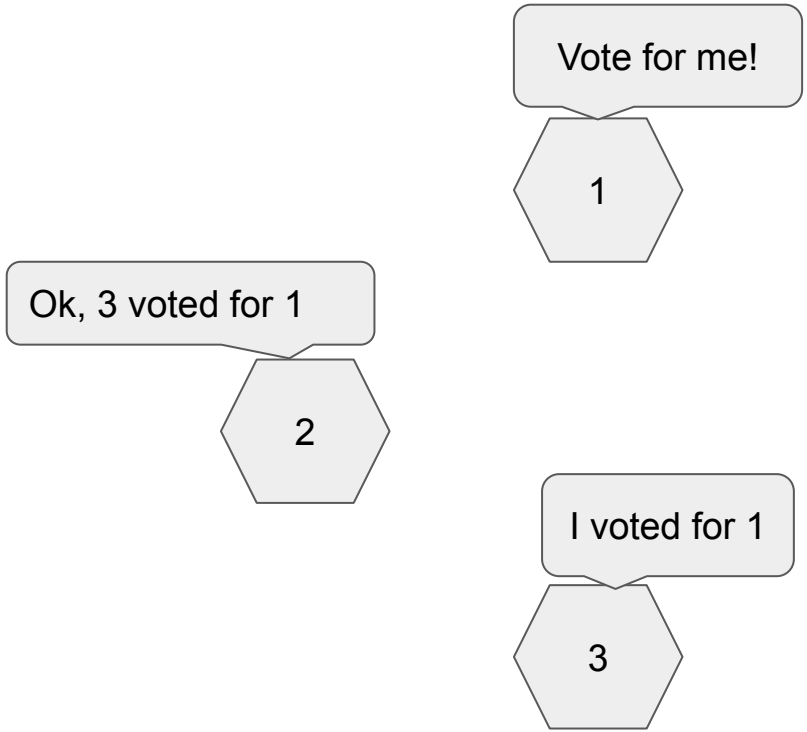


Рассылка данных от всех ко всем





Броадкаст результатов голосования



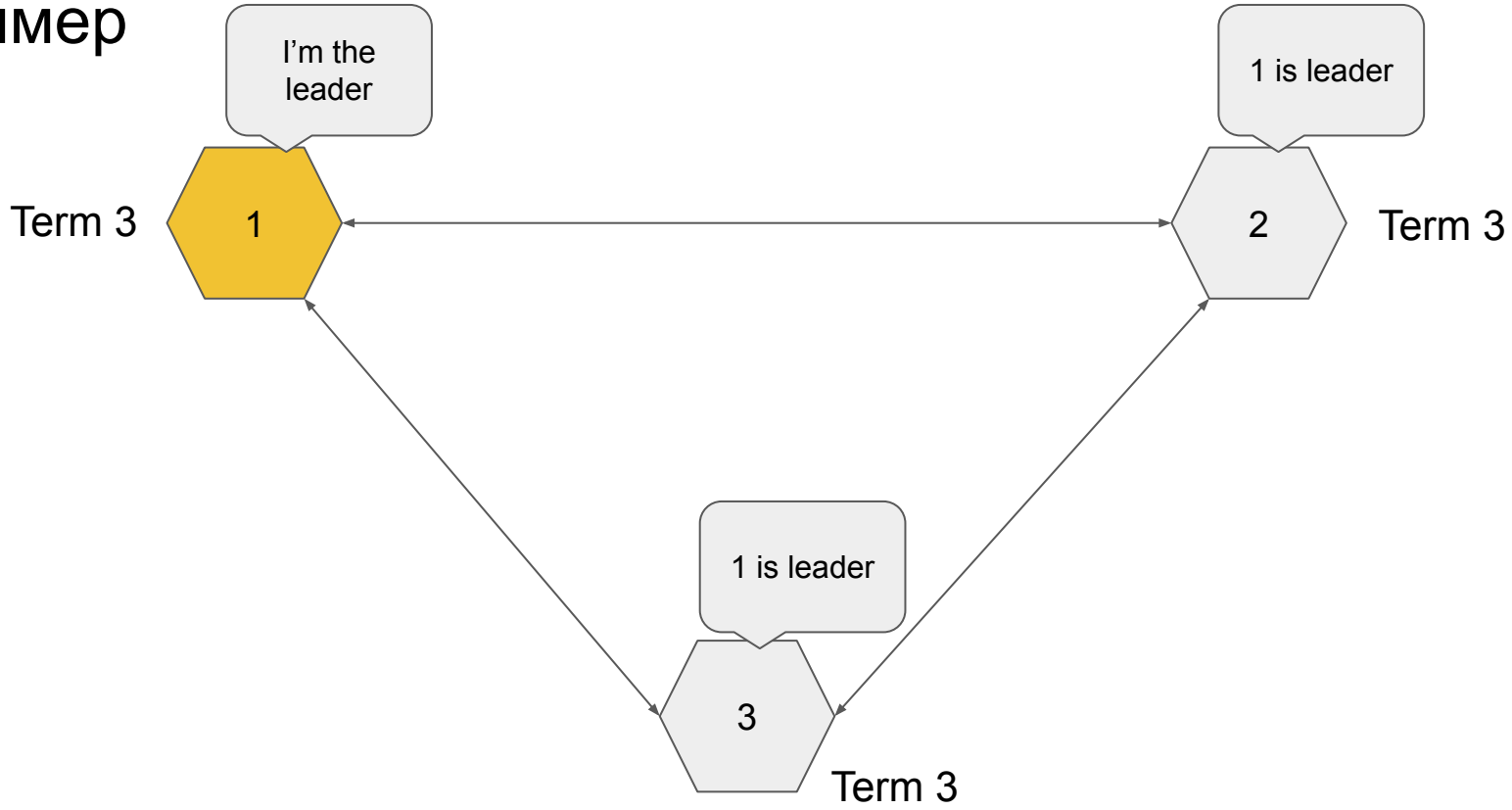
У нас был план и мы его придерживались

- ✓ Raft overview
  - Journal, Term, LSN
  - Выборы
  - Гарантии
- ✓ Raft / Tarantool: особенности
- **Надстройки Raft**
  - **Pre-Vote**
  - Split-Vote detection
  - Fencing
  - Manual promotion & Split-Brain detection

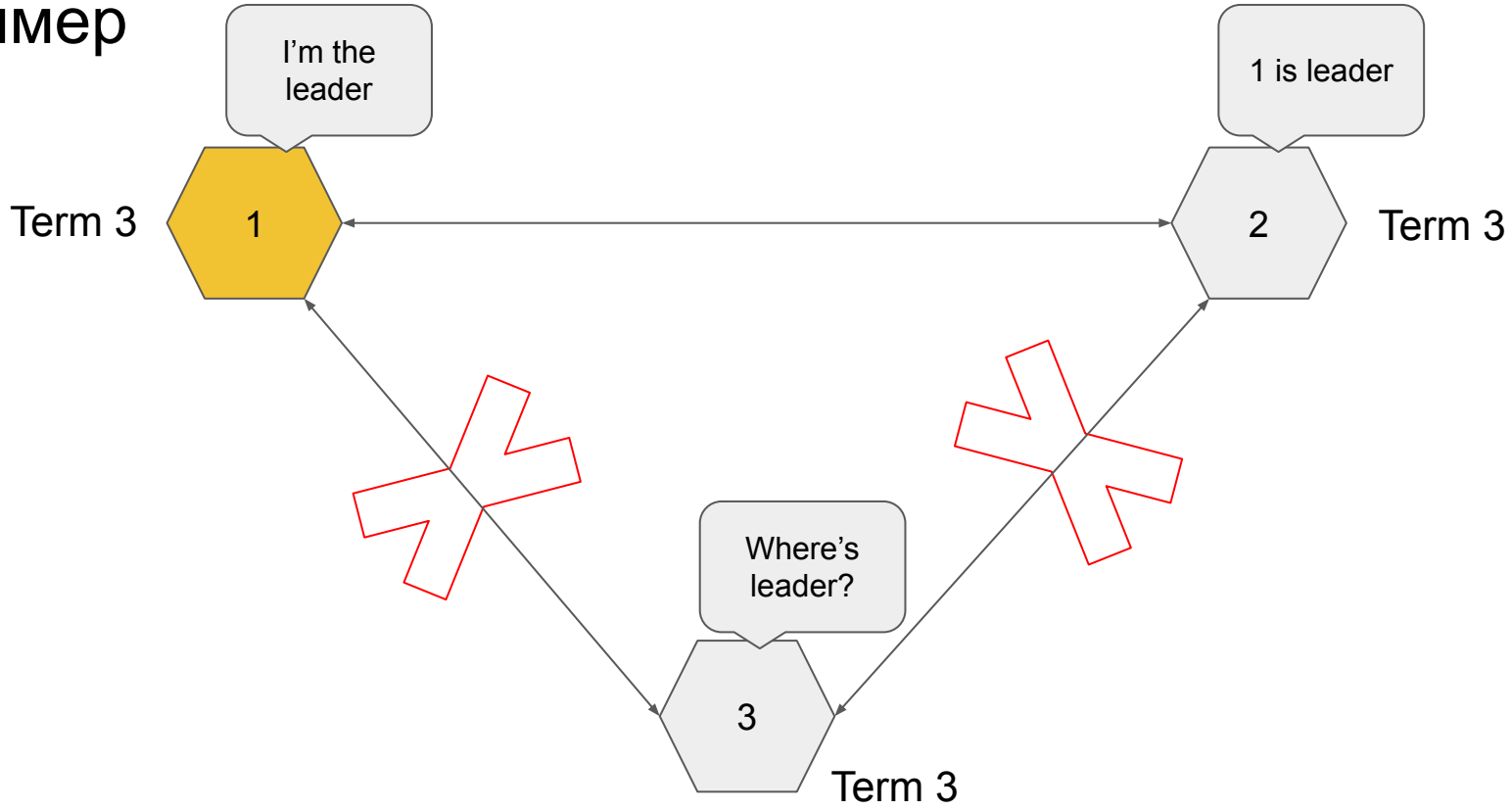
Проблема

Недоступность кластера на запись после восстановления связности

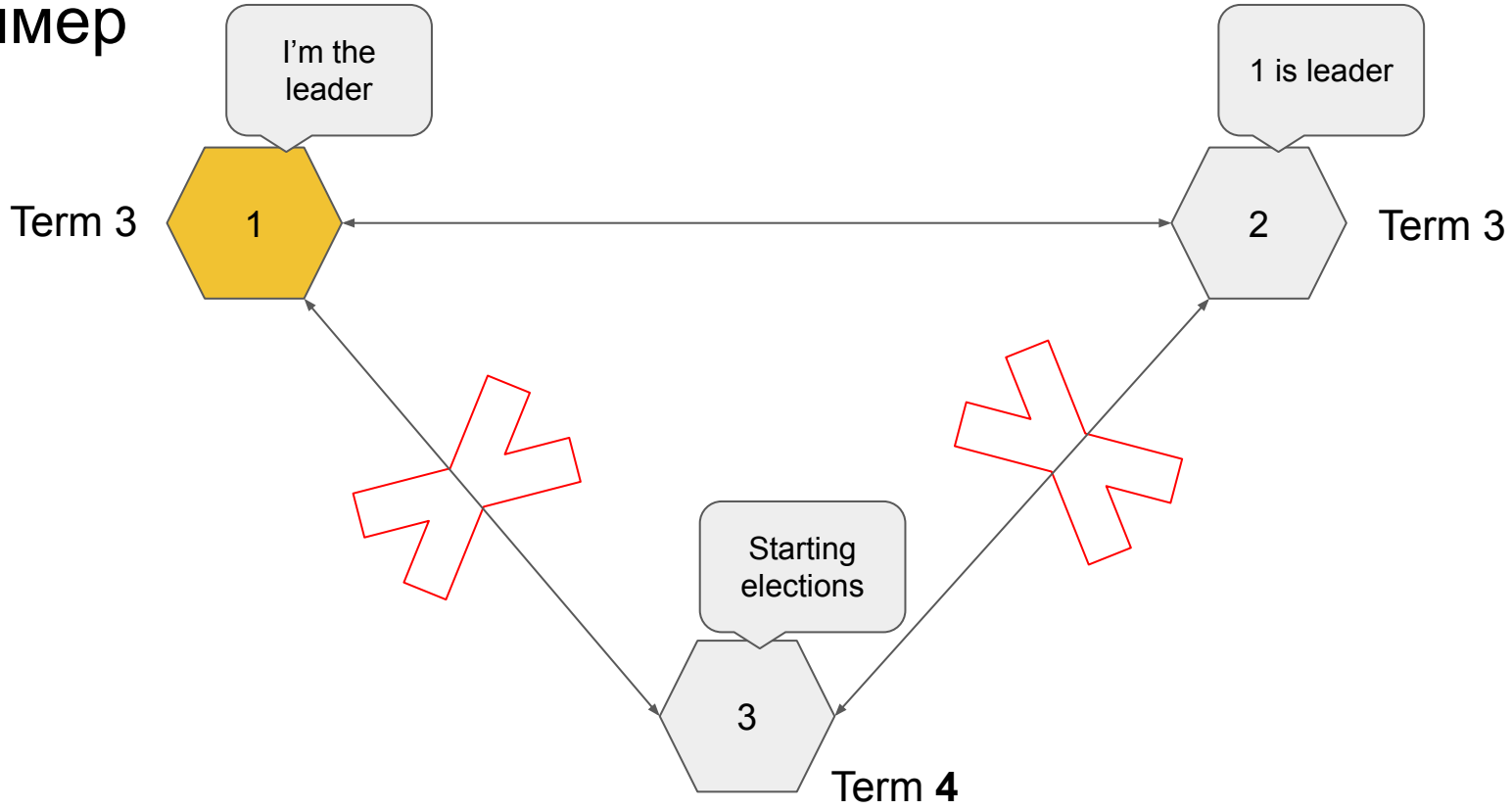
Пример



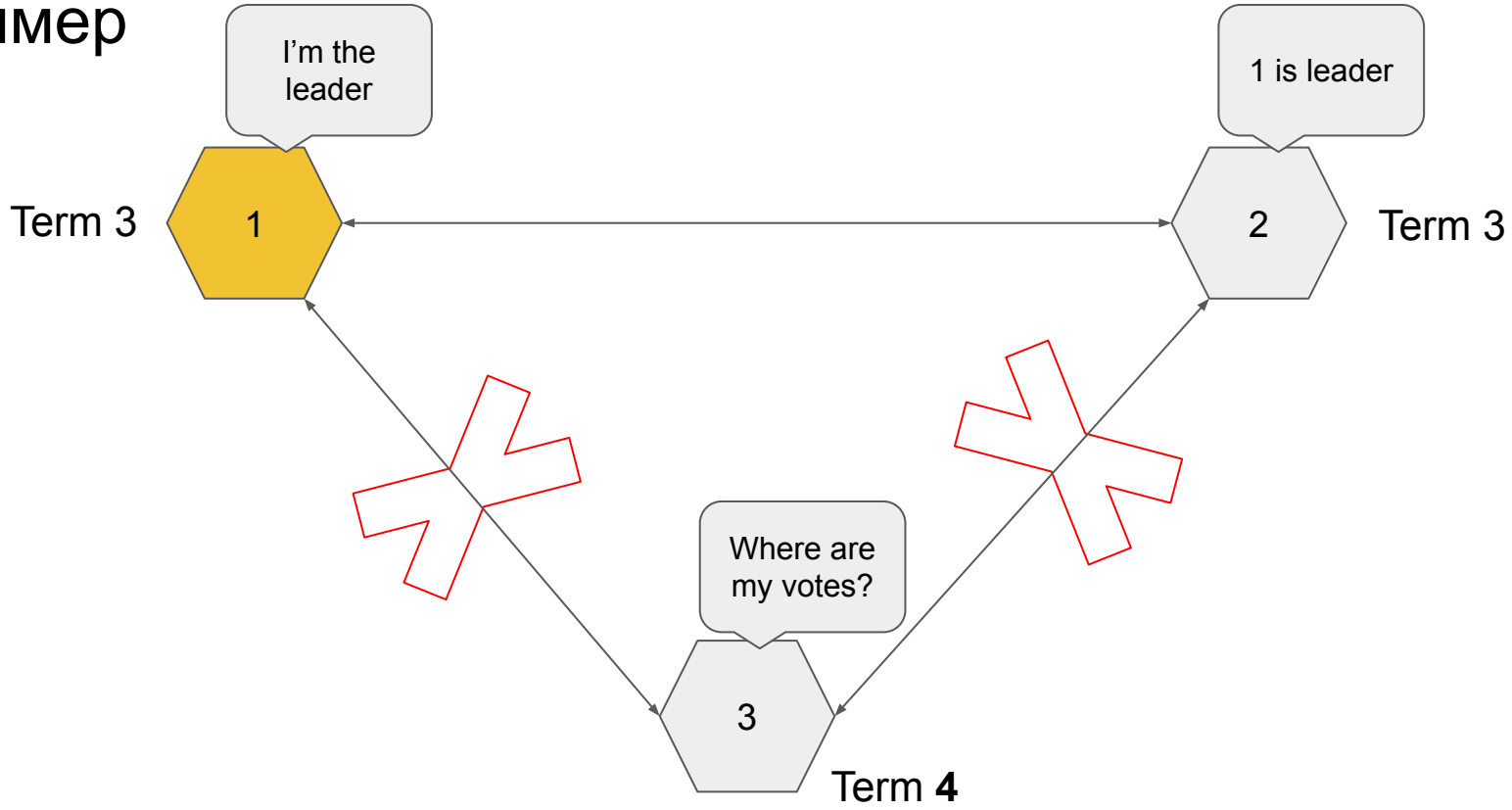
Пример


I’m the leader
1 is leader
Term 3
1
2
Term 3
Where’s leader?
3
Term 3

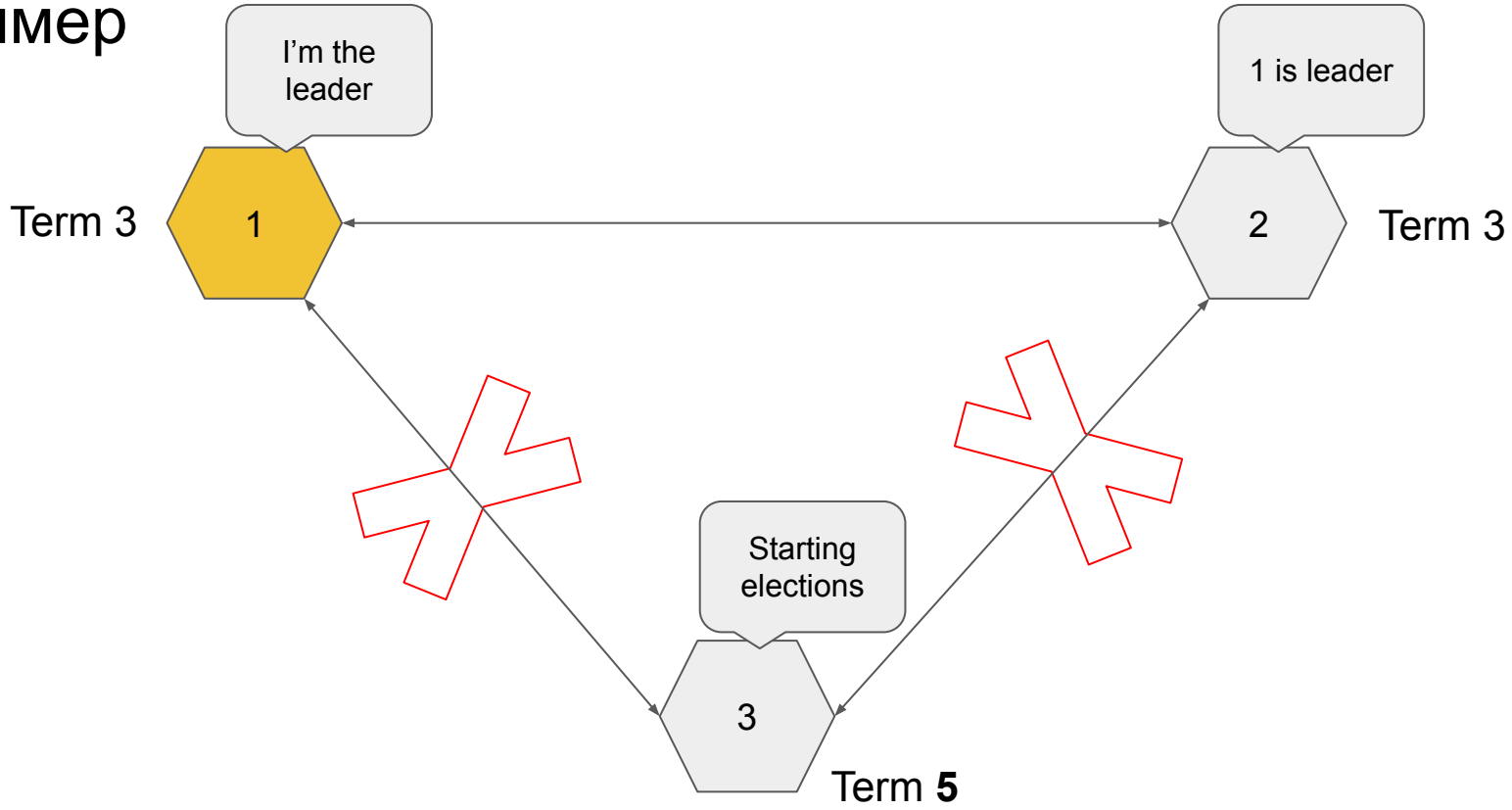
Пример


I’m the leader
1 is leader
Term 3
1
2
Term 3
Starting elections
3
Term 4

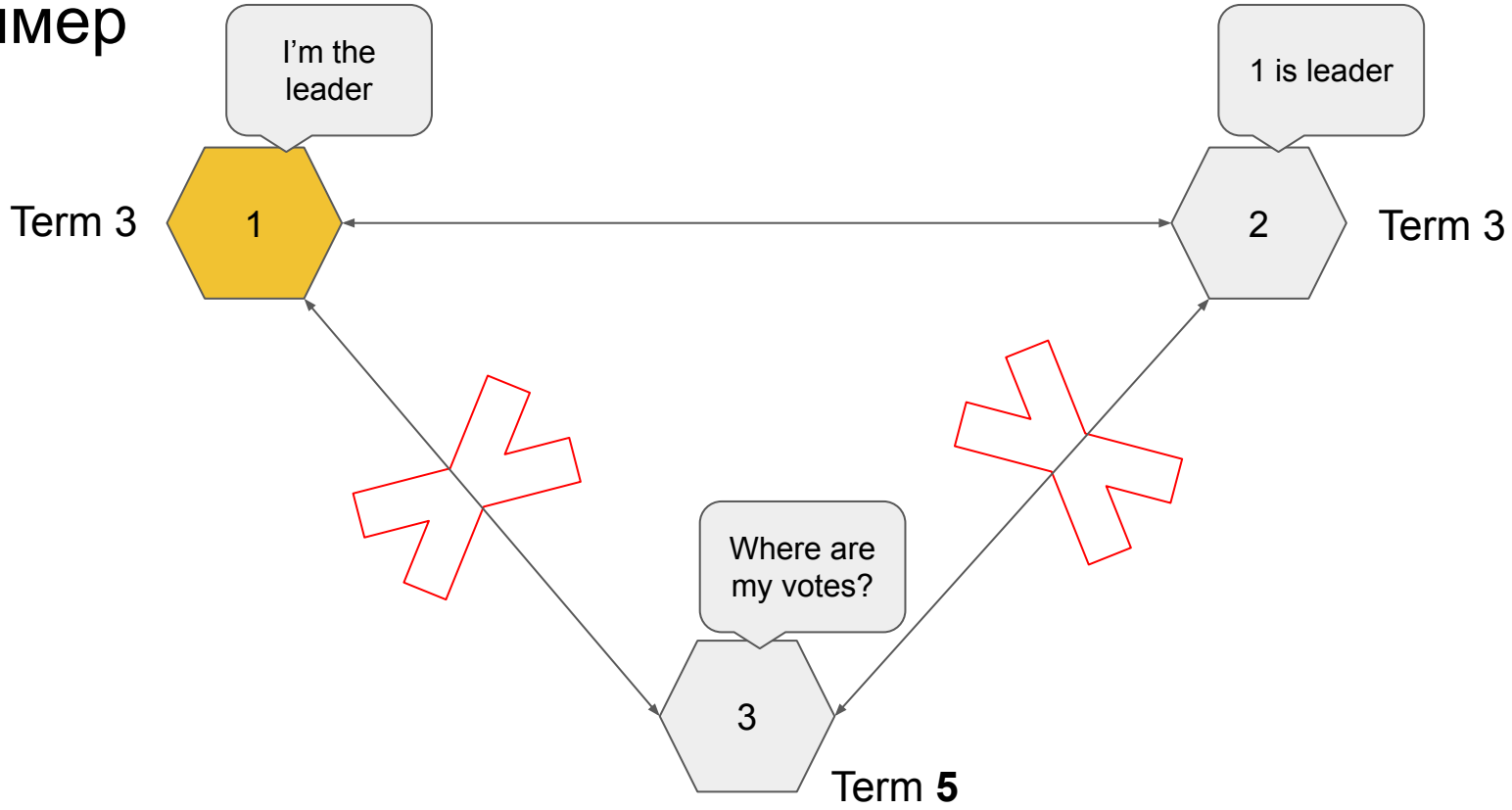
Пример


I’m the leader
1 is leader
Term 3
1
2
Term 3
Where are my votes?
3
Term 4

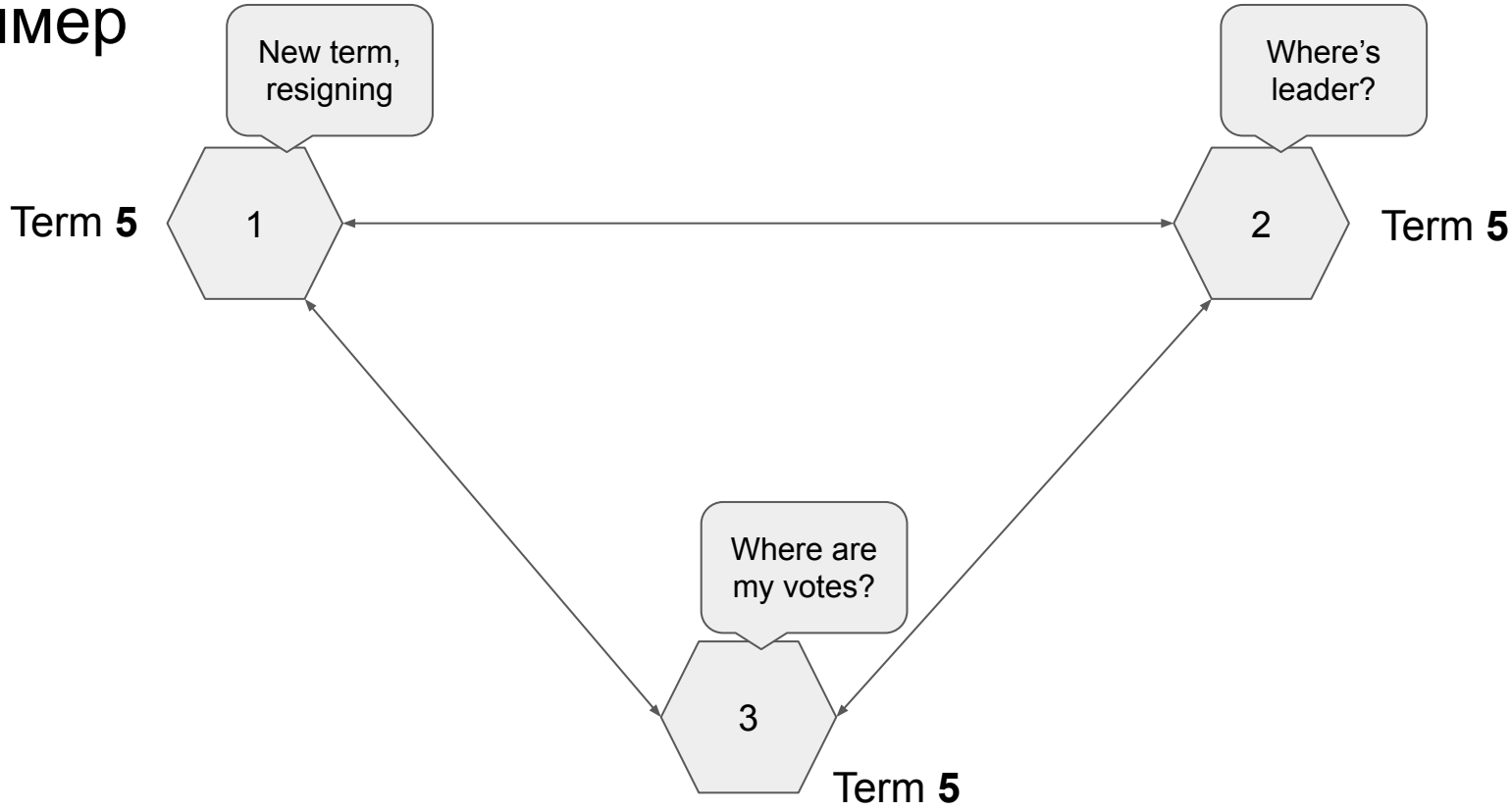
Пример


I’m the leader
1 is leader
Term 3
1
2
Term 3
Starting elections
3
Term 5

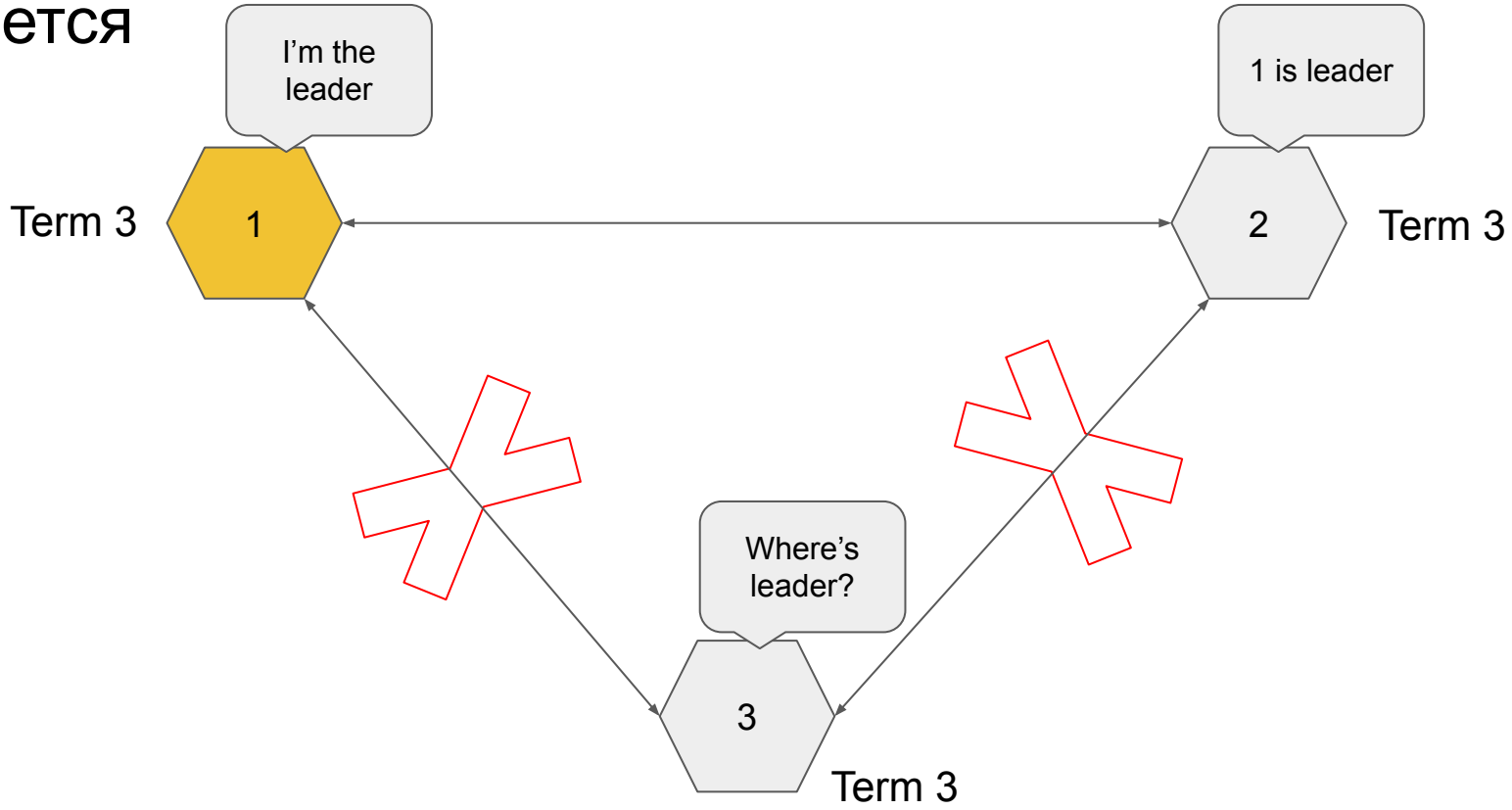
Пример


I’m the leader
1 is leader
Term 3
1
2
Term 3
Where are my votes?
3
Term 5

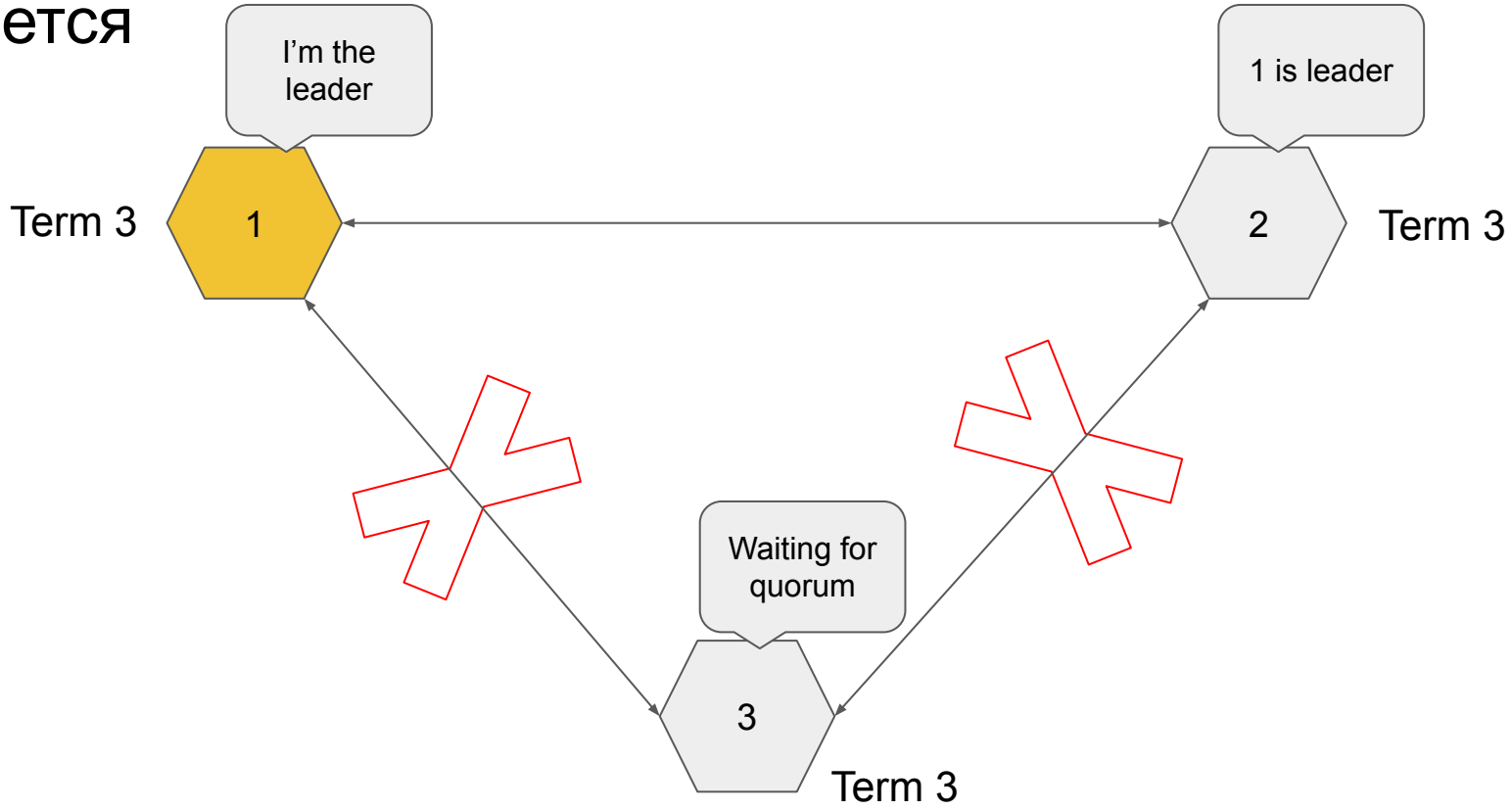
Пример


New term, resigning
Where's leader?
Term **5**
1
2
Term **5**
Where are my votes?
3
Term **5**

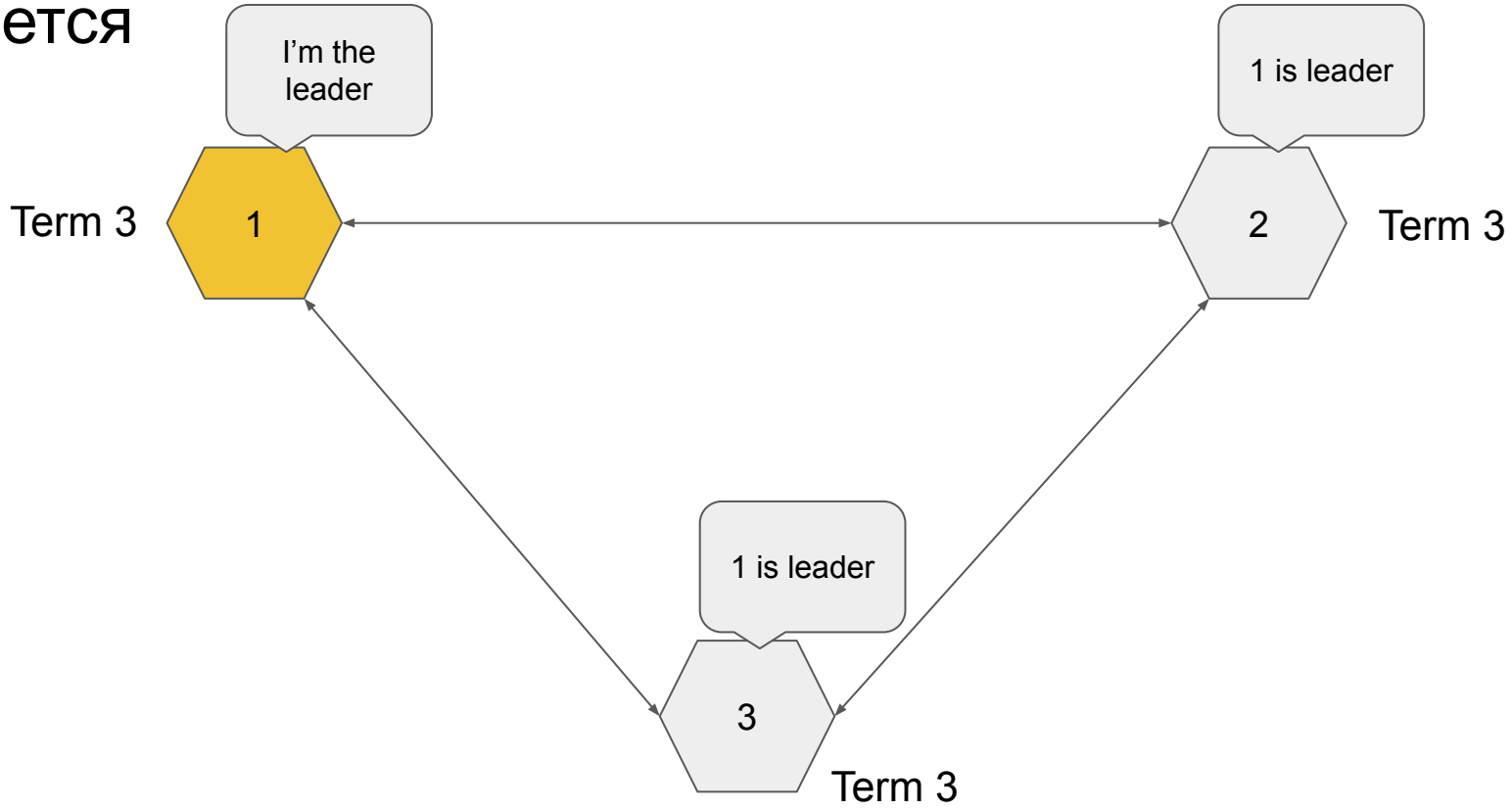
Хочется


I’m the leader
1 is leader
Term 3
1
2
Term 3
Where’s leader?
3
Term 3

Хочется


I’m the leader
1 is leader
Term 3
1
2
Term 3
Waiting for quorum
3
Term 3

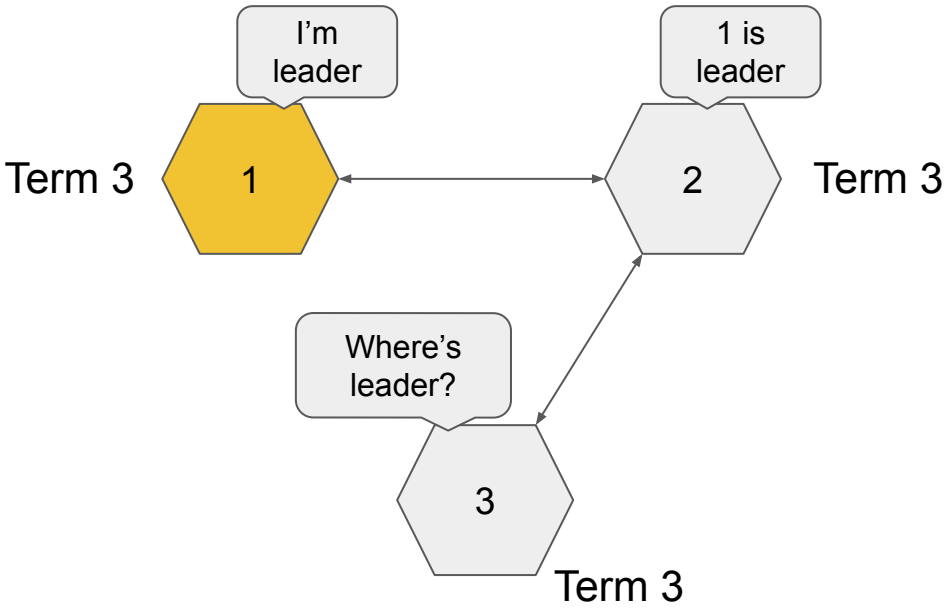
Хочется



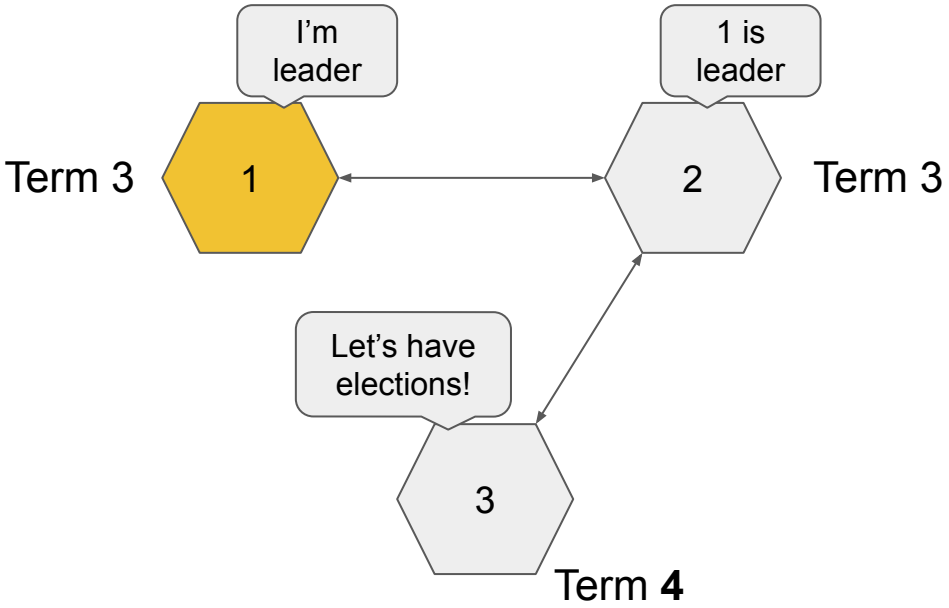
Проблема

Сервер, который не видит лидера, будет постоянно его прерывать

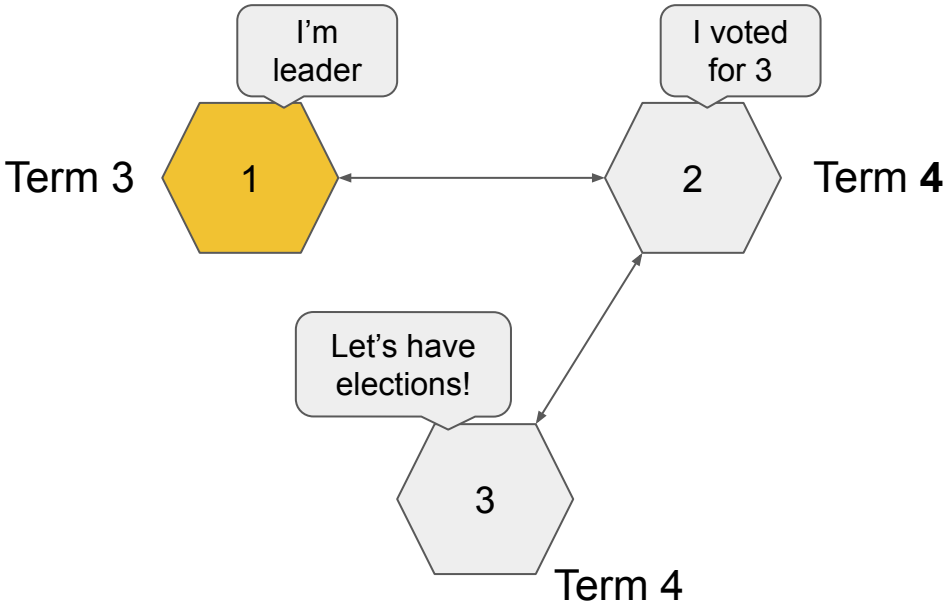
Пример



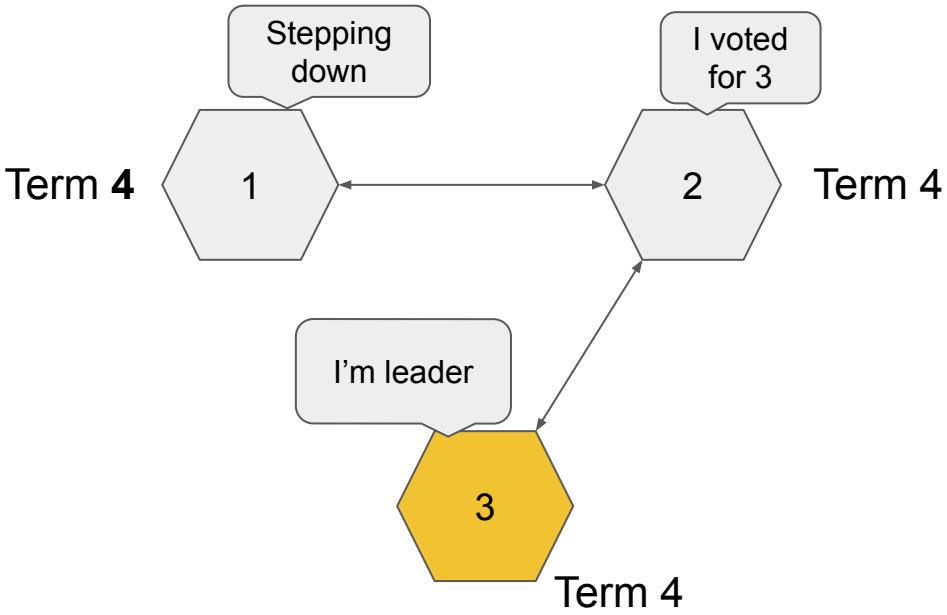
Пример


I’m leader
1 is leader
Term 3
1
2
Term 3
Let’s have elections!
3
Term 4

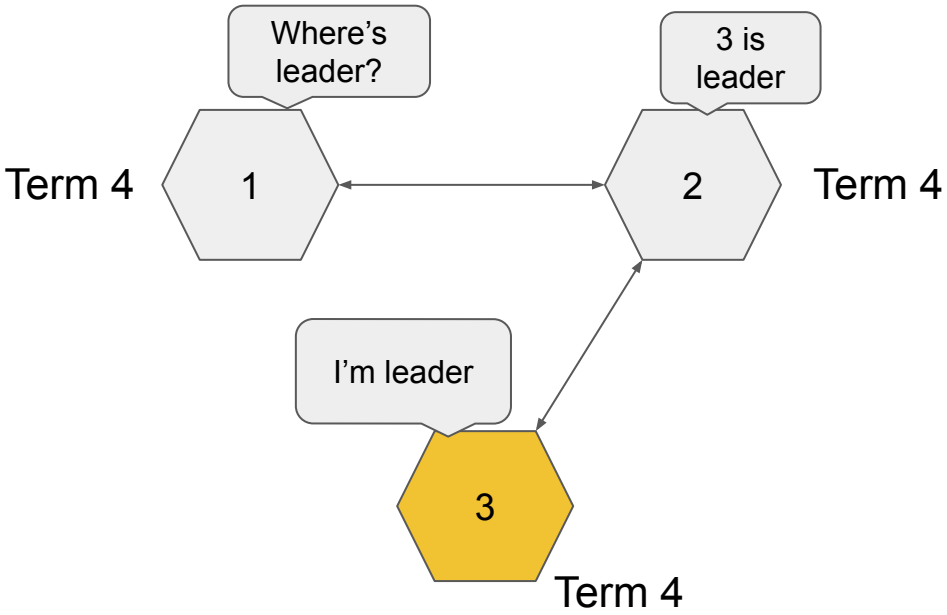
Пример


I'm leader
I voted for 3
Term 3
1
2
Term 4
Let's have elections!
3
Term 4

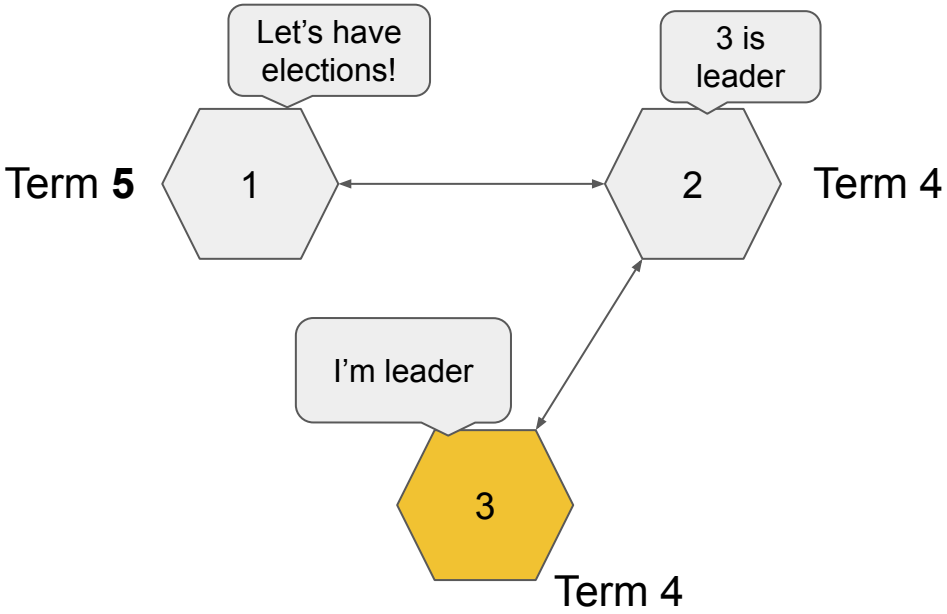
Пример



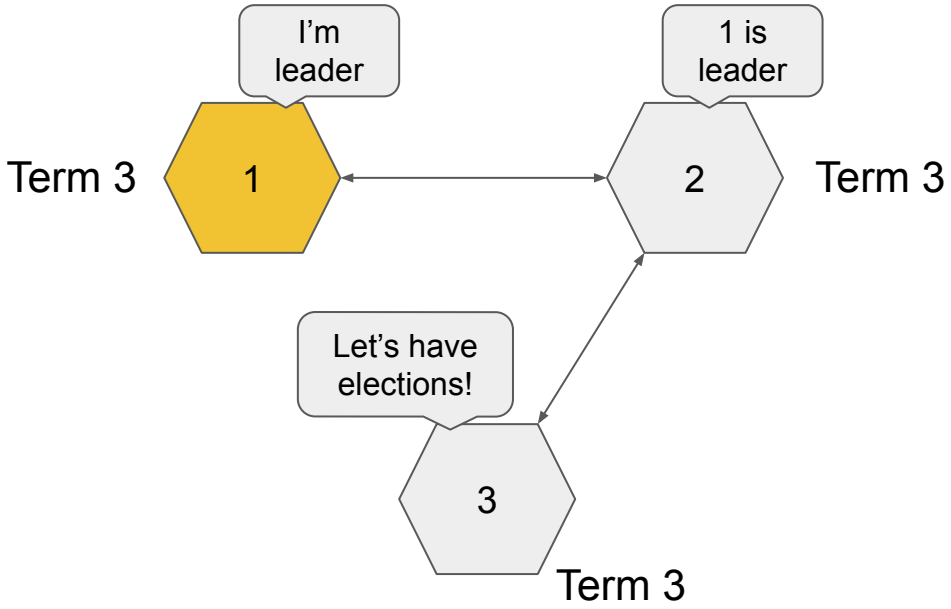
Пример



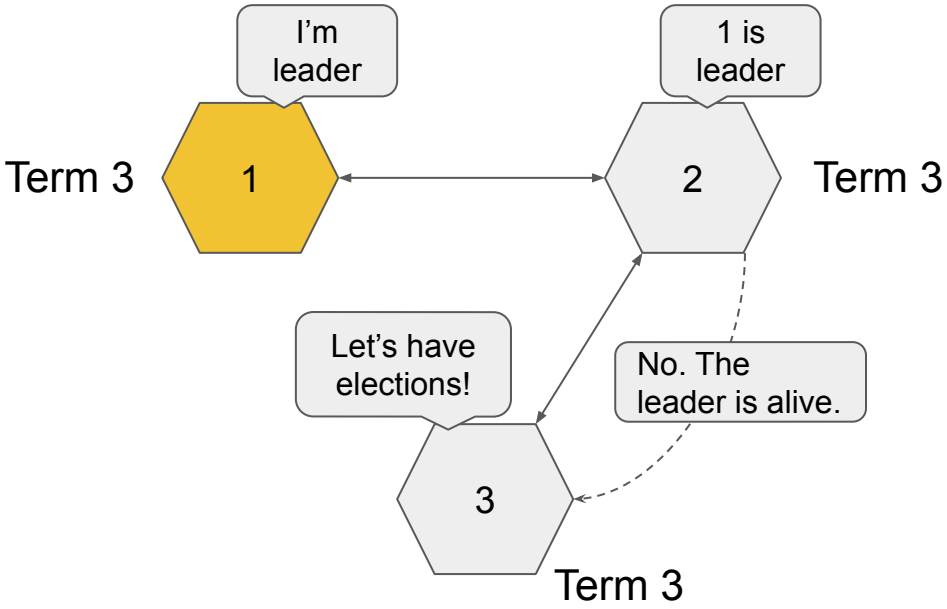
Пример


Let’s have elections!
3 is leader
Term 5
1
2
Term 4
I’m leader
3
Term 4

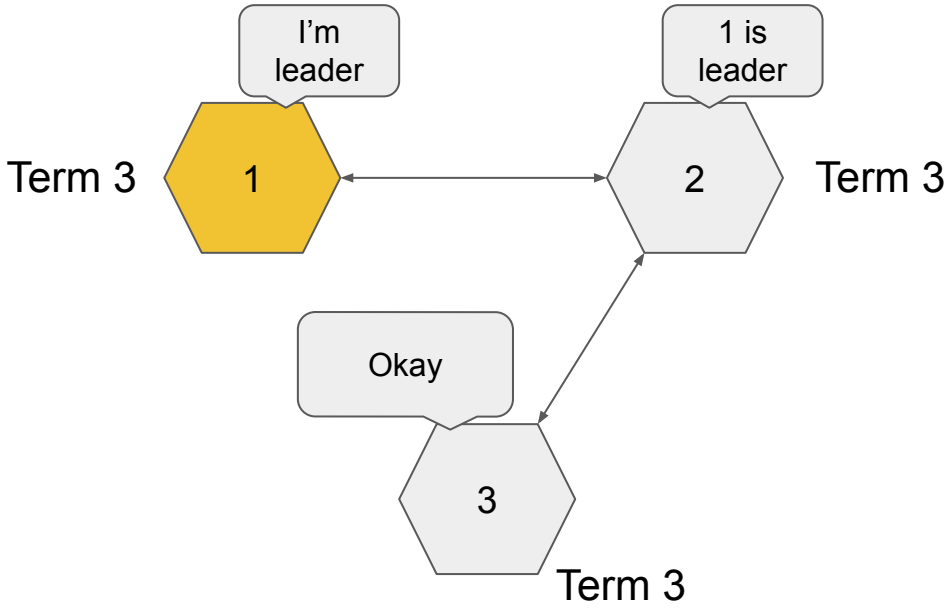
Как хочется


I’m leader
1 is leader
Term 3
1
2
Term 3
Let’s have elections!
3
Term 3

Как хочется


I’m leader
1 is leader
Term 3
1
2
Term 3
Let’s have elections!
No. The leader is alive.
3
Term 3

Как хочется



Как хочется

Итак, хотим, чтобы действующего лидера не мог прервать никто.

Варианты решения

1. Предварительные выборы
2. Игнорирование выборов голосующими, которые видят лидера
3. Pre-Vote на основе метаинформации

Варианты решения

1. **Предварительные выборы**
2. Игнорирование выборов голосующими, которые видят лидера
3. Pre-Vote на основе метаинформации

Предварительные выборы

Перед началом выборов спрашиваем всех, проголосуют ли они за нас

Предварительные выборы

Перед началом выборов спрашиваем всех, проголосуют ли они за нас

Ответ положительный, если голосующий сам не видит лидера, и если проголосовал бы за кандидата в обычных выборах

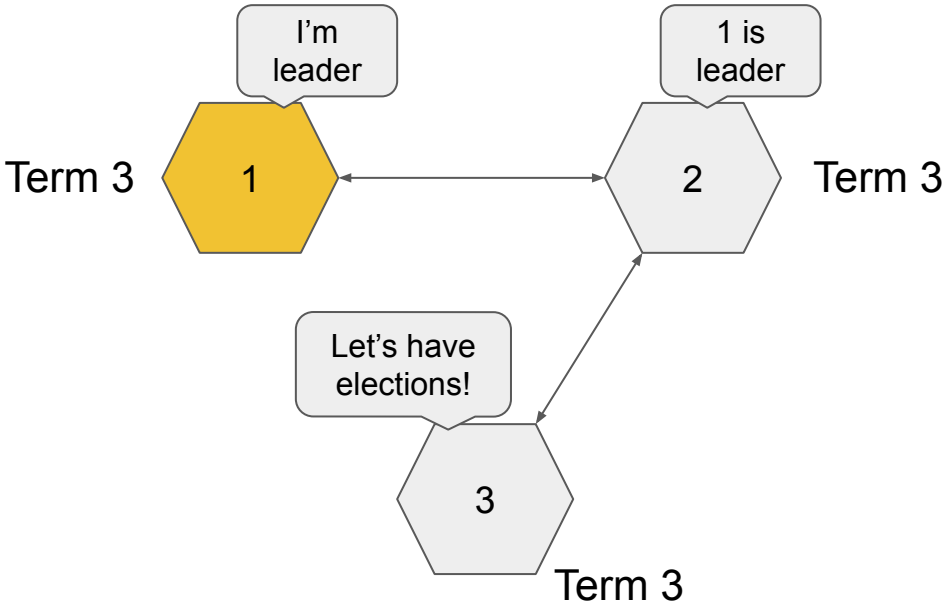
Предварительные выборы

Перед началом выборов спрашиваем всех, проголосуют ли они за нас

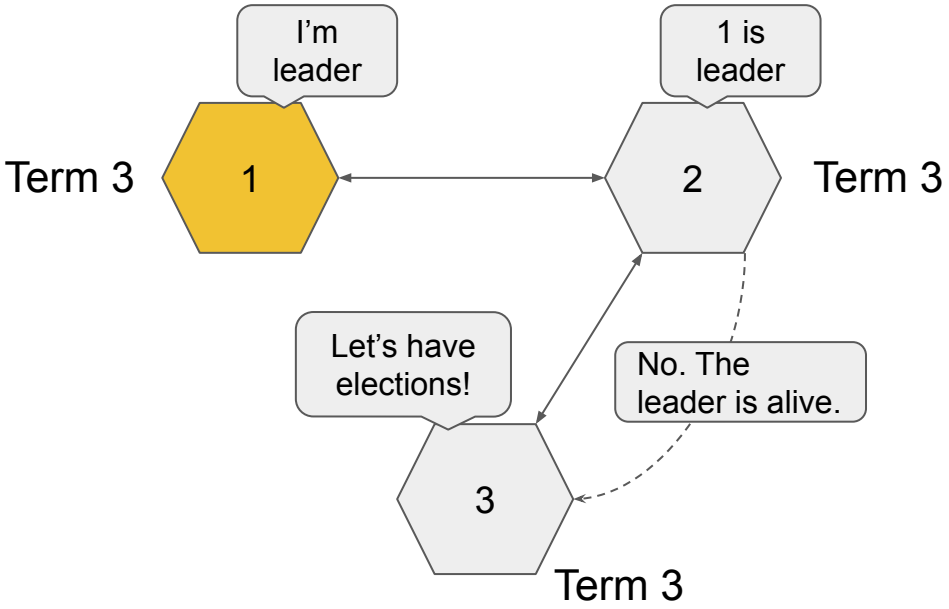
Ответ положительный, если голосующий сам не видит лидера, и если проголосовал бы за кандидата в обычных выборах

После получения кворума положительных ответов начинаем выборы

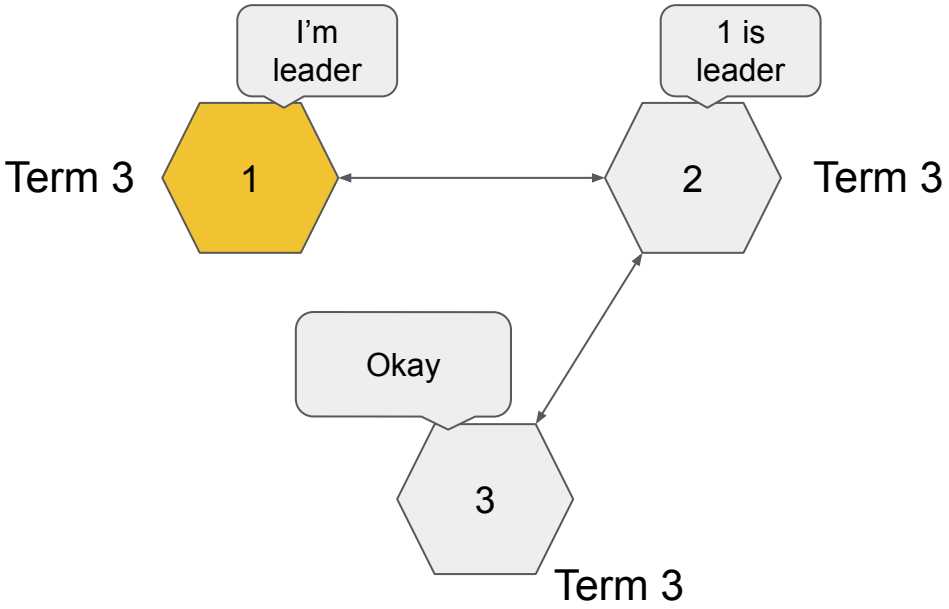
Почему это помогает

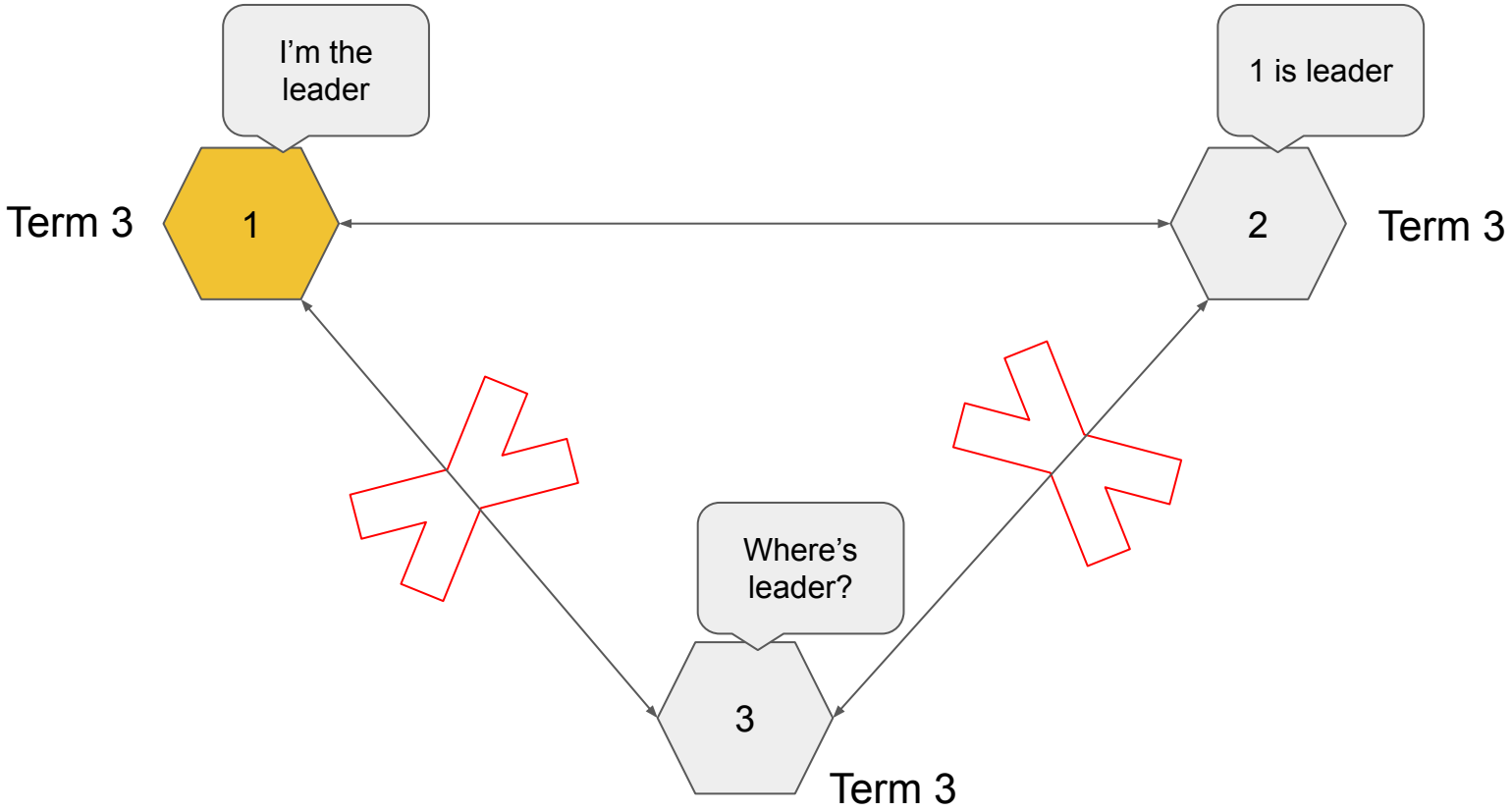


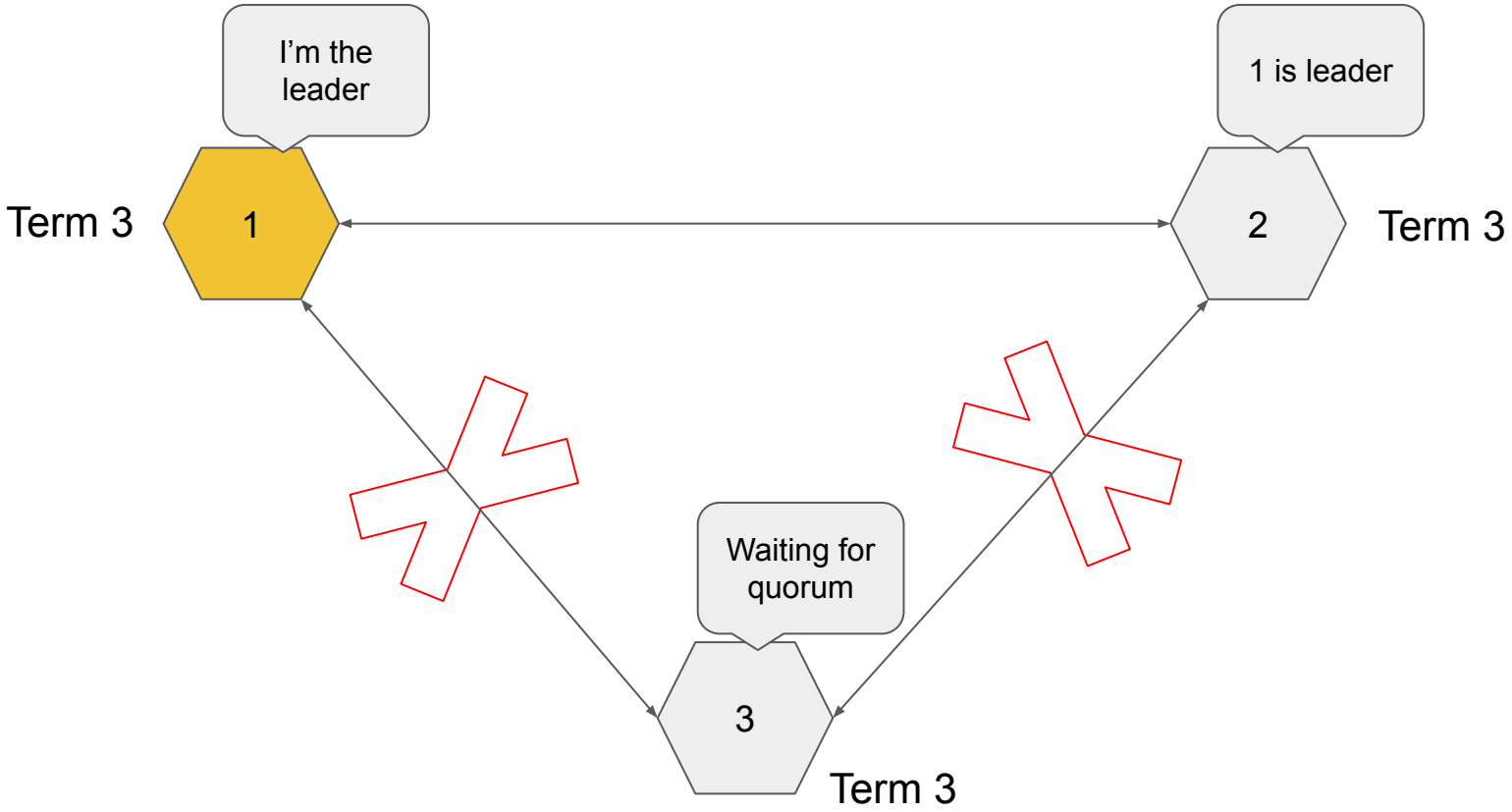
Почему это помогает



Почему это помогает




I’m the leader
1 is leader
Term 3
1
2
Term 3
Where’s leader?
3
Term 3


I’m the leader
1 is leader
Term 3
1
2
Term 3
Waiting for quorum
3
Term 3

Предварительные выборы

Недостаток - расширение протокола, ломающее обратную совместимость с существующими инсталляциями

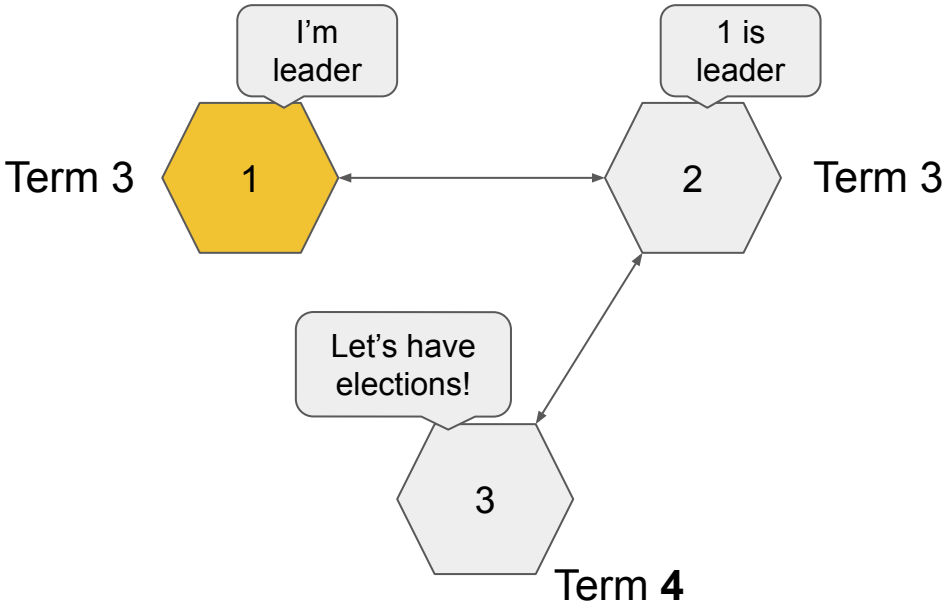
Варианты решения

1. Предварительные выборы
2. **Игнорирование выборов голосующими, которые видят лидера**
3. Pre-Vote на основе метаинформации

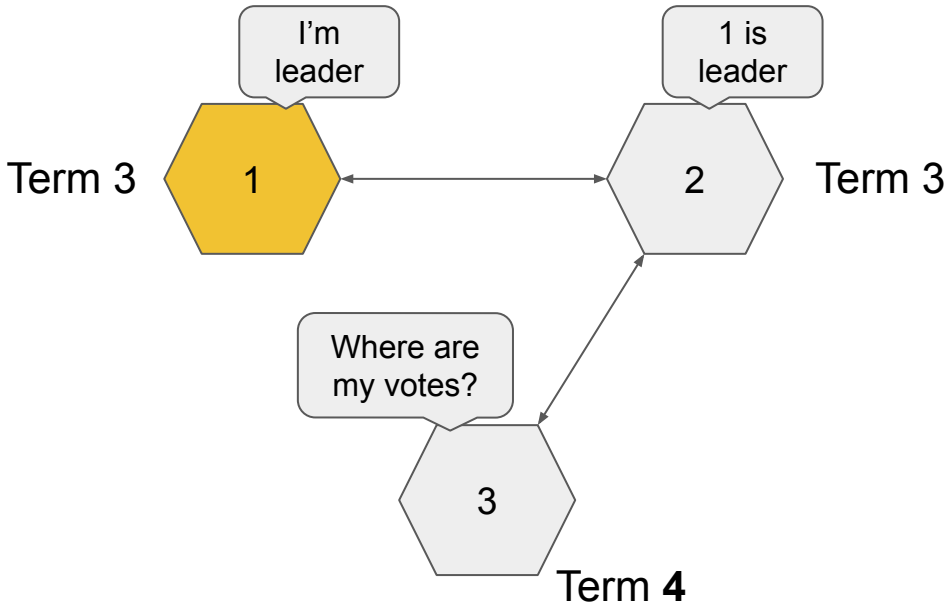
Игнорирование выборов

Сервер, видящий лидера напрямую, игнорирует любые запросы на голос от соседей

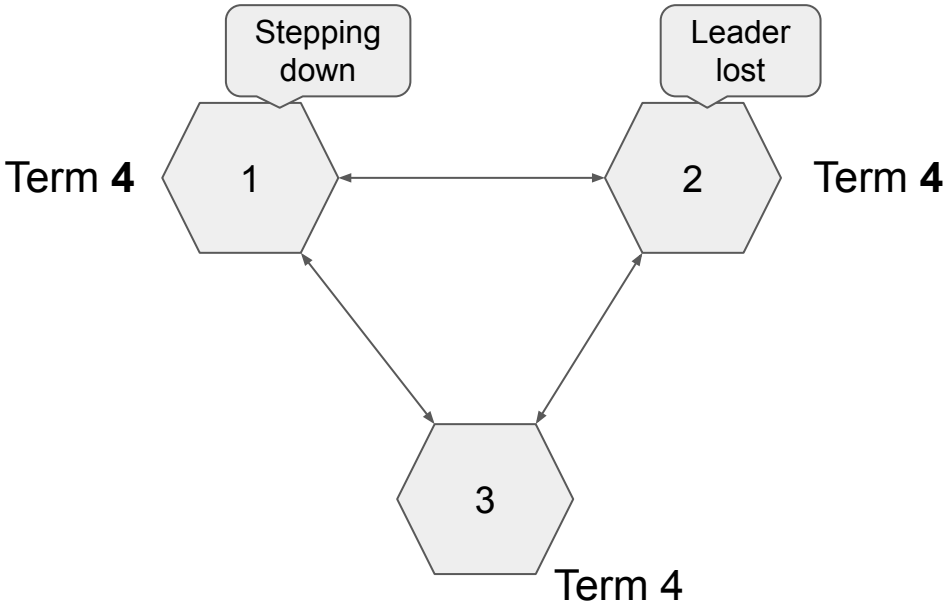
Почему это помогает



Почему это помогает



Почему это **не** помогает



Игнорирование выборов

Достоинство – не вносим изменения в протокол

Недостаток – решает только часть проблемы

Варианты решения

1. Предварительные выборы
2. Игнорирование выборов голосующими, которые видят лидера
3. **Pre-Vote на основе метаинформации**

Pre-Vote на основе метаинформации

В Tarantool каждый сервер знает состояние Raft соседей

Pre-Vote на основе метаинформации

В Tarantool каждый сервер знает состояние Raft соседей

Добавим к рассылаемому состоянию информацию о том, виден ли лидер

Pre-Vote на основе метаинформации

Узлу известно

1. К скольким из соседей он подключен
2. Кто из соседей видит лидера

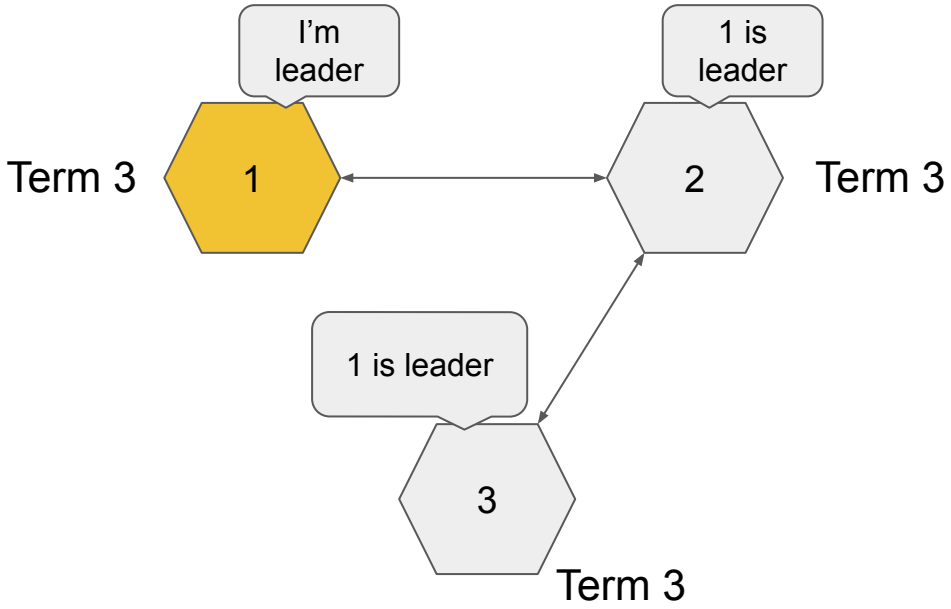
Pre-Vote на основе метаинформации

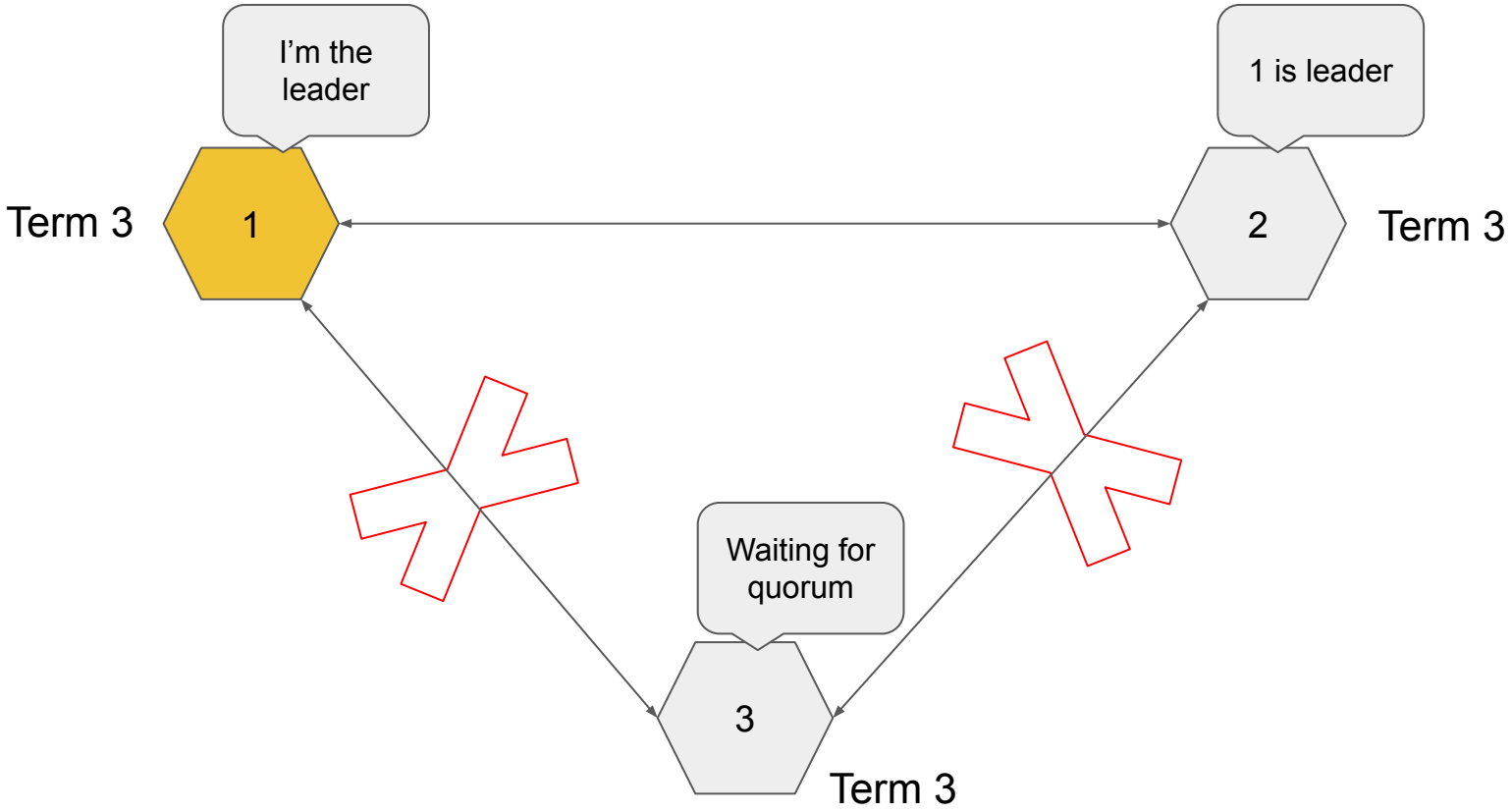
Узлу известно

1. К скольким из соседей он подключен
2. Кто из соседей видит лидера

Кто-то видит лидера в текущем терме? Не начинаем выборы

Почему это помогает




I’m the leader
1 is leader
Term 3
1
2
Term 3
Waiting for quorum
3
Term 3

Pre-Vote на основе метаинформации

Достоинство – добиваемся тех же результатов, что и с предварительной стадией выборов, ценой меньшего расширения протокола

Модификация #2: Split-Vote detection

Проблема

Долгая недоступность кластера на запись после потери лидера

Проблема

Долгая недоступность кластера на запись после потери лидера

Каждый раунд выборов, в результате которого не выбран лидер – ≥ 5 секунд простоя

Проблема

Долгая недоступность кластера на запись после потери лидера

Каждый раунд выборов, в результате которого не выбран лидер –
≥ 5 секунд простоя

В случае успешных выборов лидер появляется в самом начале раунда

Проблема

Долгая недоступность кластера на запись после потери лидера

Каждый раунд выборов, в результате которого не выбран лидер –
≥ 5 секунд простоя

В случае успешных выборов лидер появляется в самом начале раунда

Если голоса разделились, оставшееся время раунда – простой

Проблема

Решение

У нас есть широковещательная рассылка отданных голосов

Решение

У нас есть широковещательная рассылка отданных голосов

Каждый из узлов ведёт счёт голосования

Решение

У нас есть широковещательная рассылка отданных голосов

Каждый из узлов ведёт счёт голосования

Стало понятно, что выборы никто не выиграет? Новый раунд

Решение

У нас есть широковещательная рассылка отданных голосов

Каждый из узлов ведёт счёт голосования

Стало понятно, что выборы никто не выиграет? Новый раунд

Экономия до 90% времени на каждый раунд с разделившимися голосами

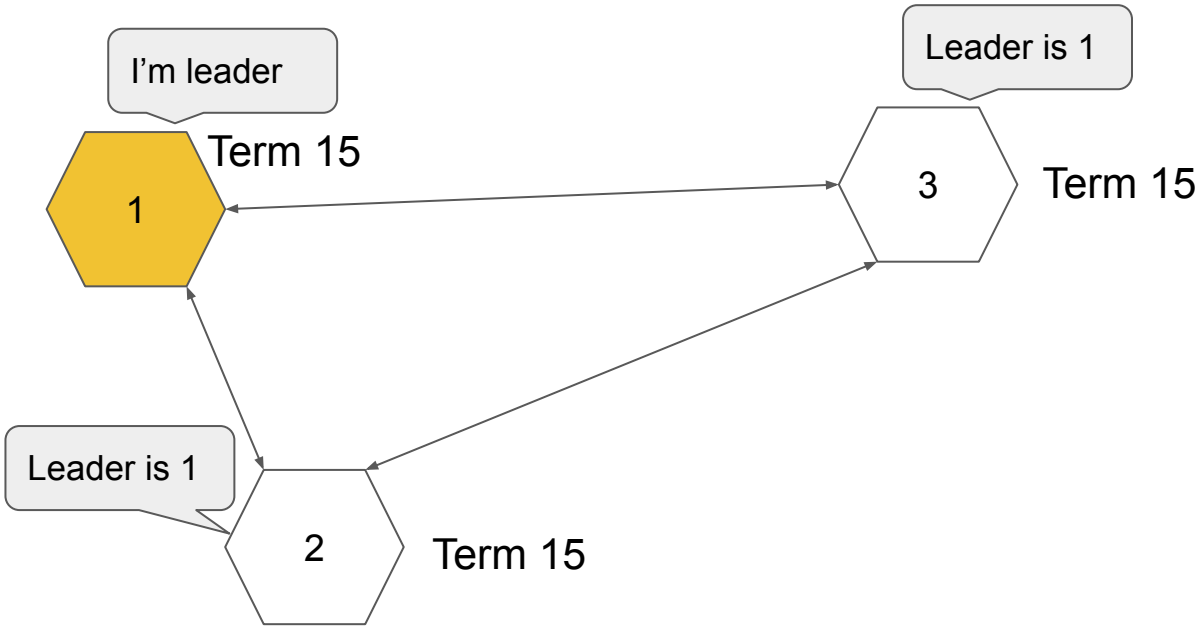
У нас был план и мы его придерживались

- ✓ Raft overview
  - Journal, Term, LSN
  - Выборы
  - Гарантии
- ✓ Raft / Tarantool: особенности
- **Надстройки Raft**
  - ✓ Pre-Vote
  - ✓ Split-Vote detection
  - **Fencing**
  - Manual promotion & Split-Brain detection

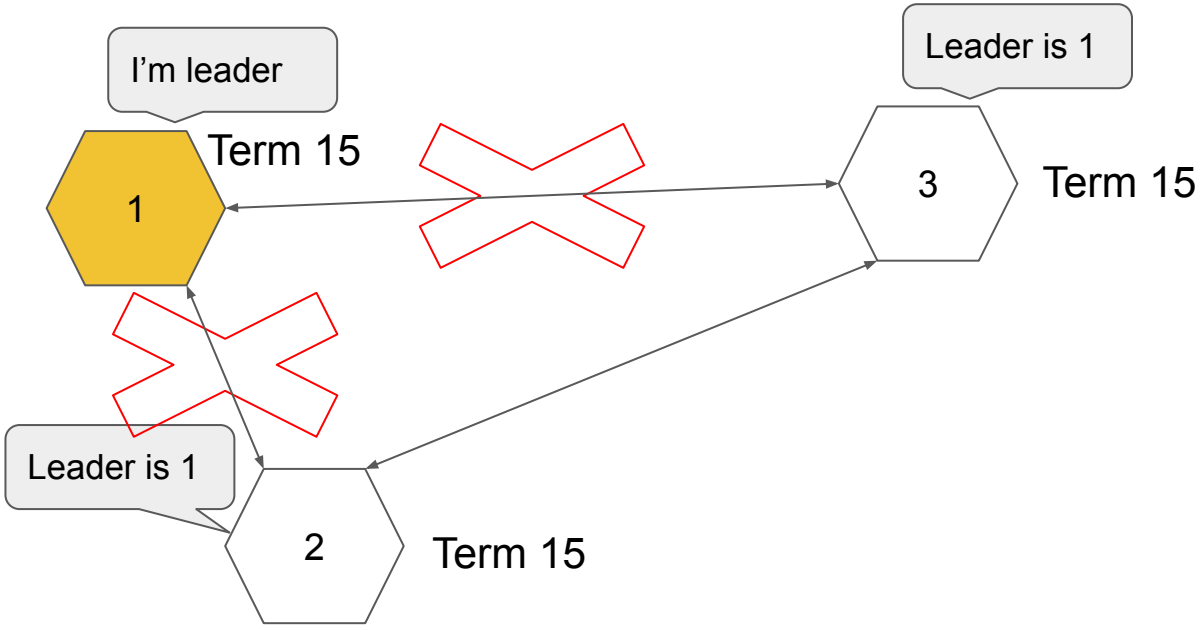
Проблема

- Для записи в кластер необходимо обратиться к текущему лидеру
- Возможность существования нескольких нод, считающих себя лидерами одновременно (в разных term)
- Не более одного из “лидеров” сможет на самом деле произвести синхронную запись

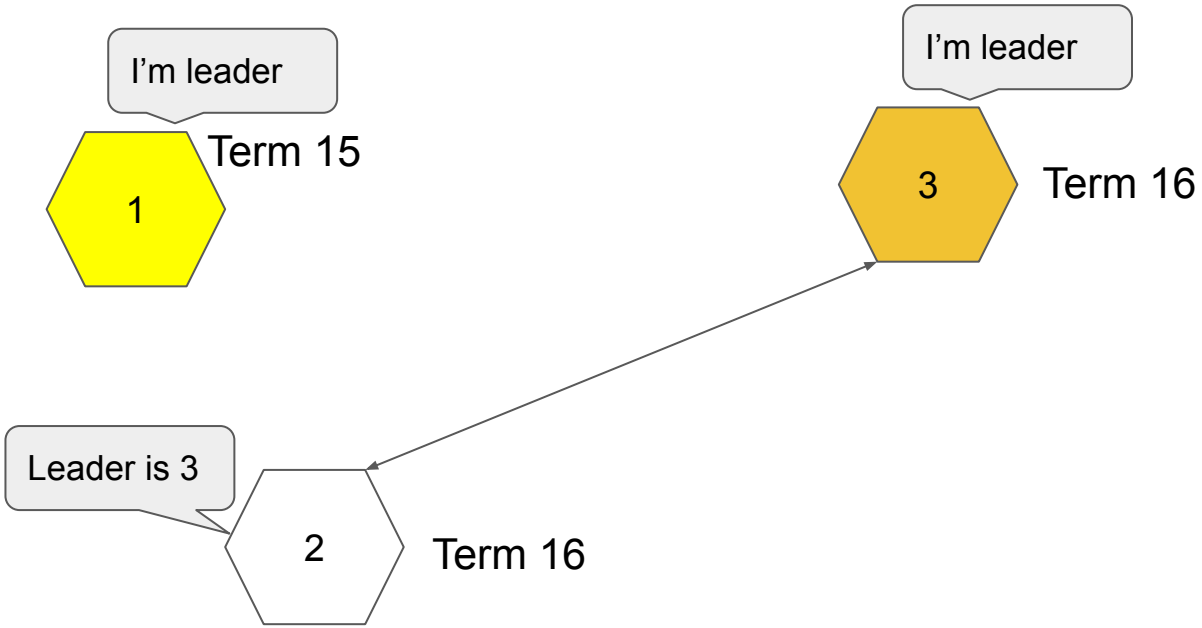
Проблема



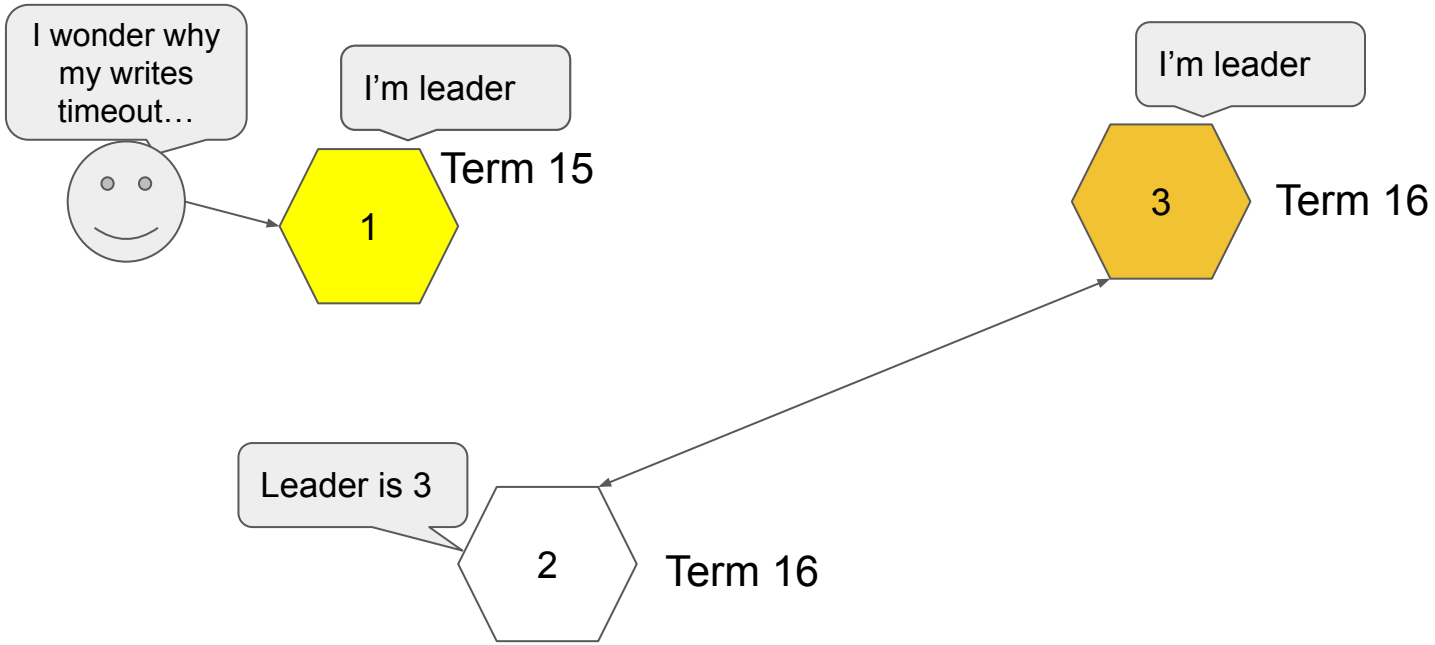
Проблема



Проблема



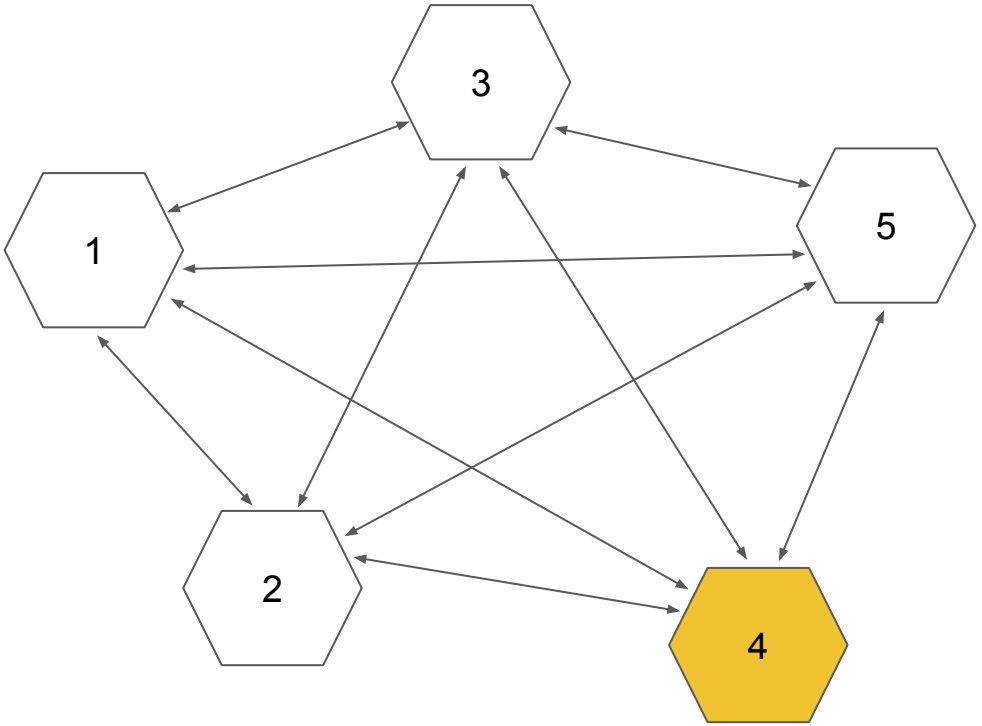
Проблема


I wonder why my writes timeout…
I’m leader
Term 15
1
I’m leader
3
Term 16
Leader is 3
2
Term 16

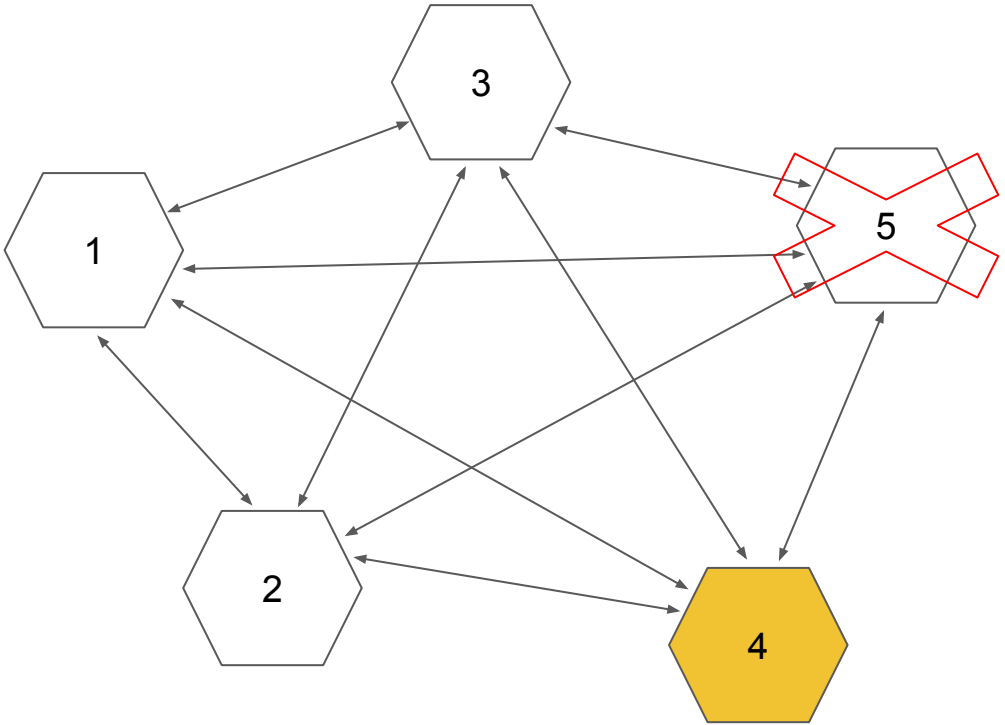
Проблема

- При наличии pre-vote можно получить “заблокированный” кластер

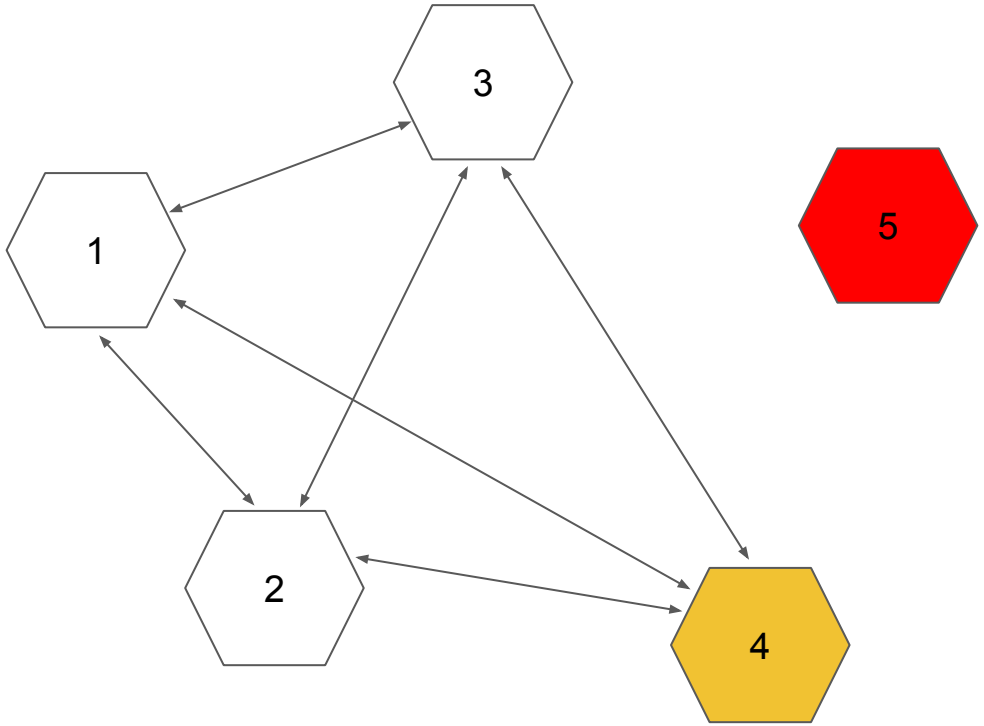
Проблема



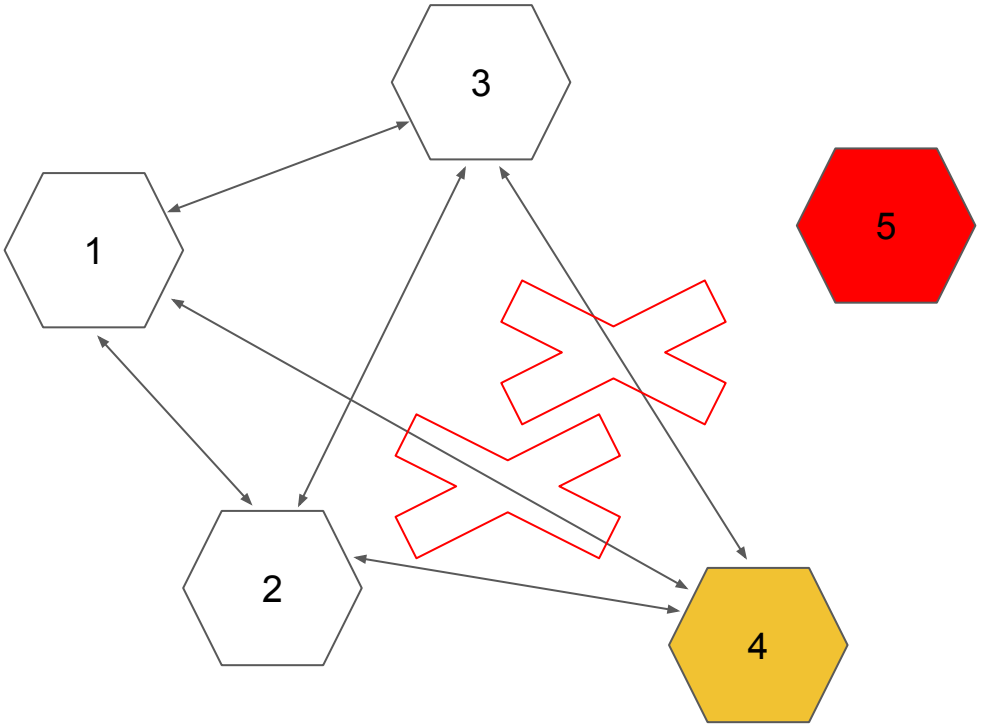
Проблема



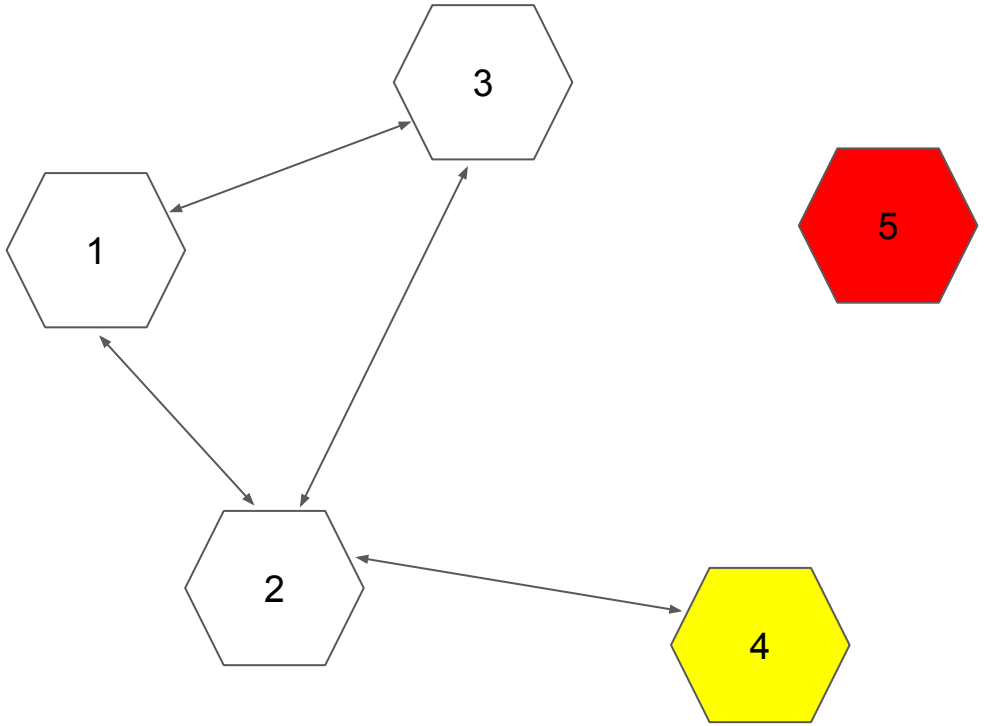
Проблема



Проблема



Проблема



Способы решения

- **Для определения текущего лидера обращаться ко всем нодам кластера, запрашивая текущий терм и состояние ноды (лидер или нет)**
- Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей
- Доработать RAFT в Tarantool

Спросить всех

- Не решает проблему с “заблокированного” кластера
- Для внешней системы увеличивается время в течении которого она не может производить запись
- Требуется реализация данного решения во всех внешних системах

Способы решения

- Для определения текущего лидера обращаться ко всем нодам кластера, запрашивая текущий терм и состояние ноды (лидер или нет)
- **Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей**
- Доработать RAFT в Tarantool

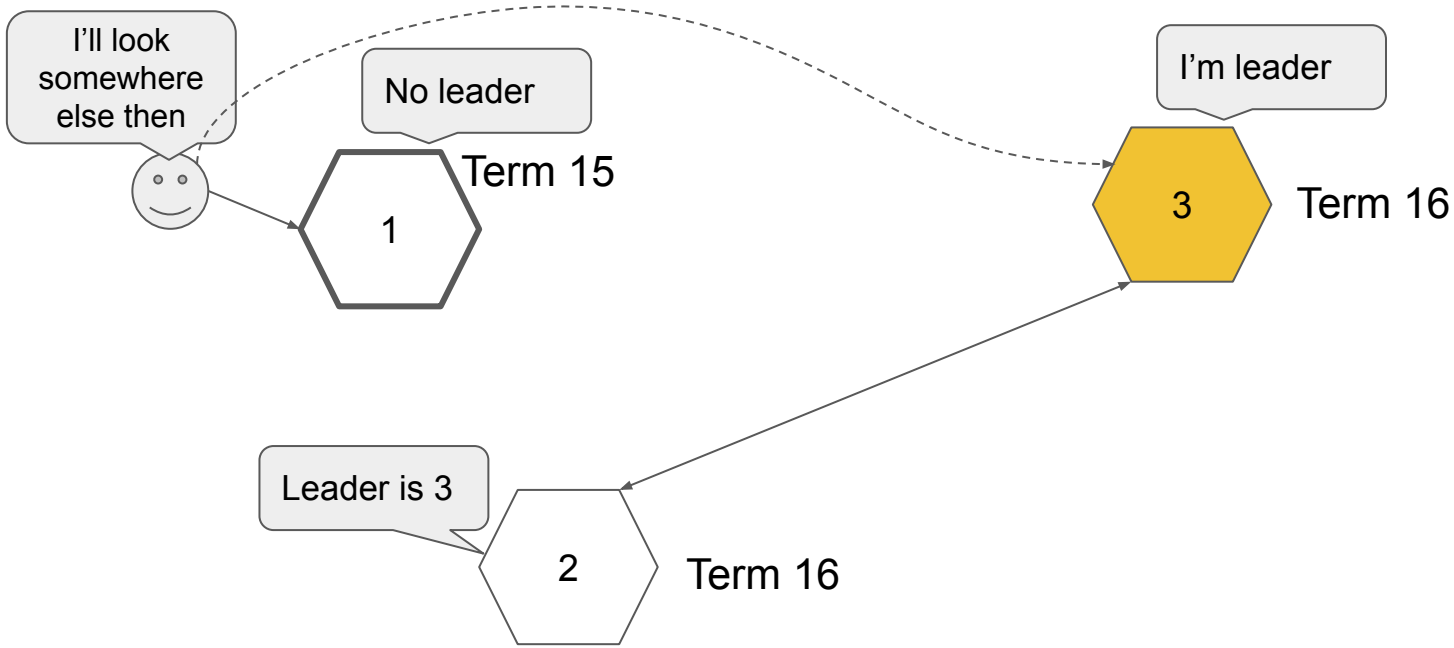
Внешнее назначение

- Доверяем решение проблемы третьей системе
- Еще одна точка отказа

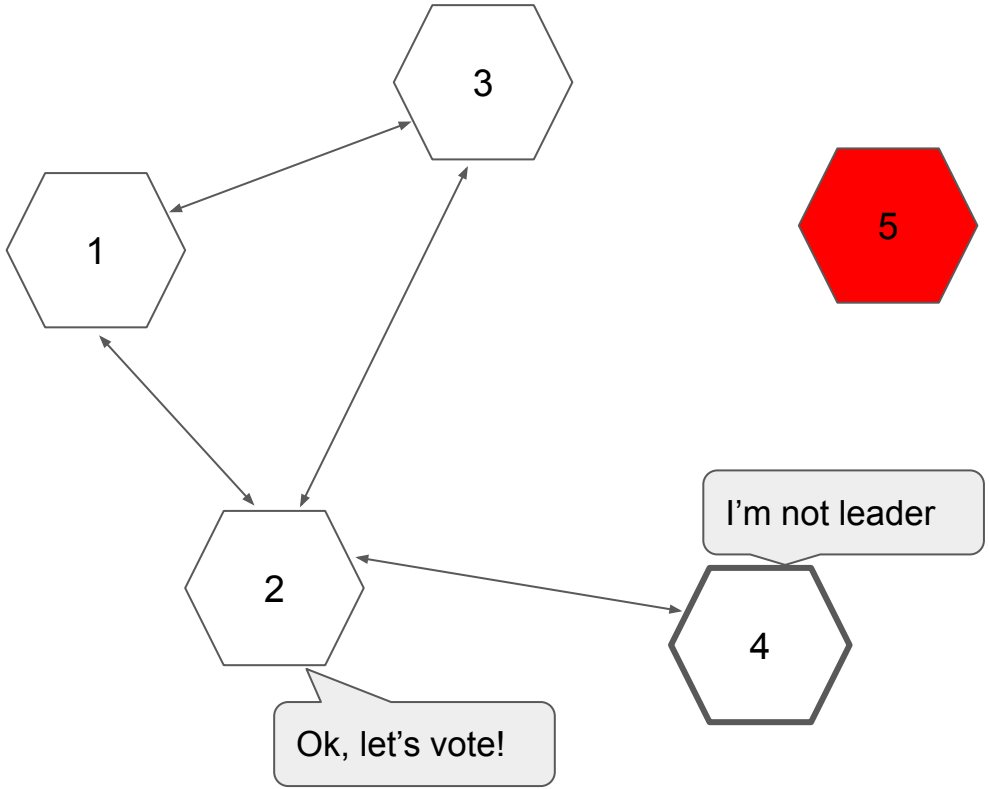
Способы решения

- Для определения текущего лидера обращаться ко всем нодам кластера, запрашивая текущий терм и состояние ноды (лидер или нет)
- Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей
- **Доработать RAFT в Tarantool**

Как хочется


I’ll look somewhere else then
No leader
Term 15
1
I’m leader
3
Term 16
Leader is 3
2
Term 16

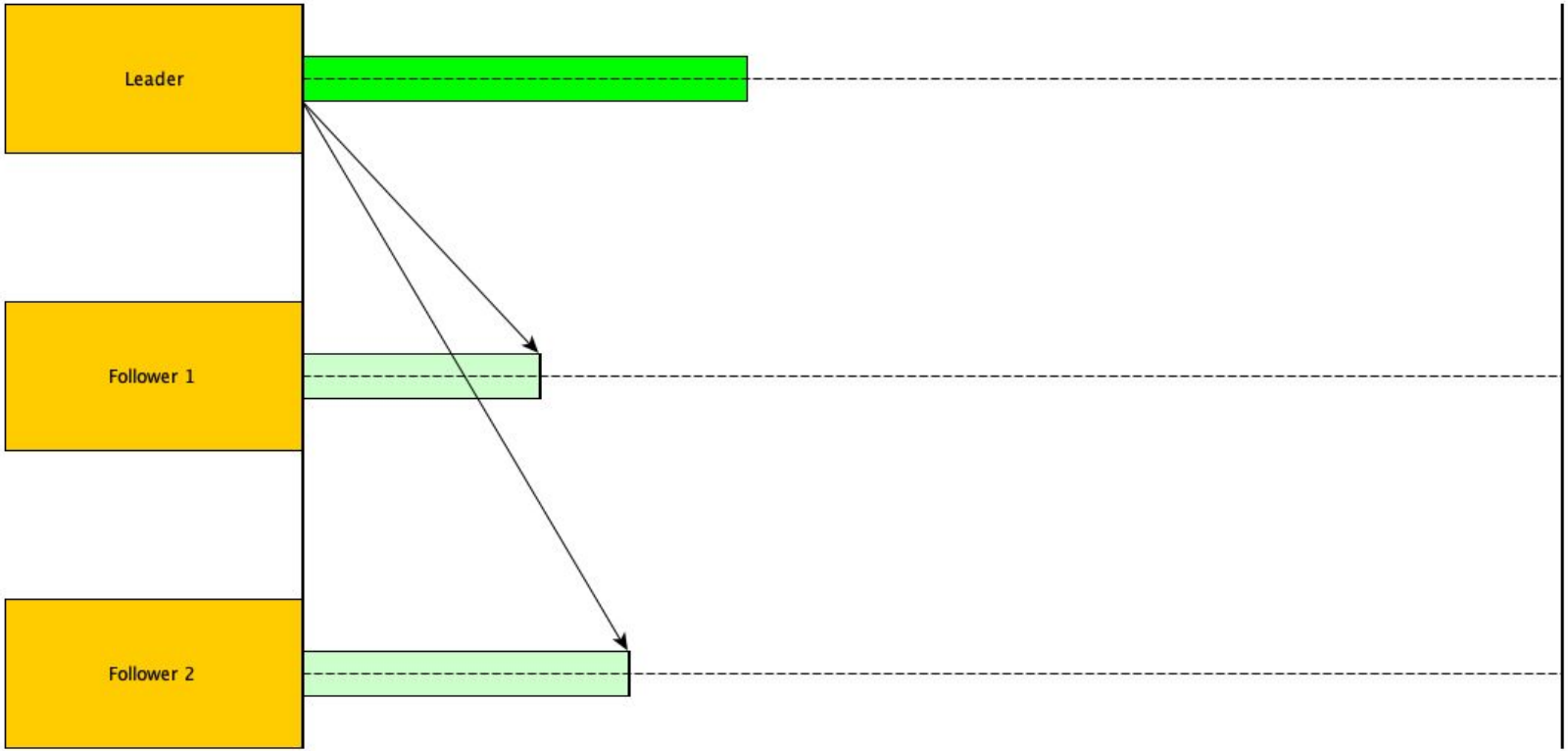
Как хочется



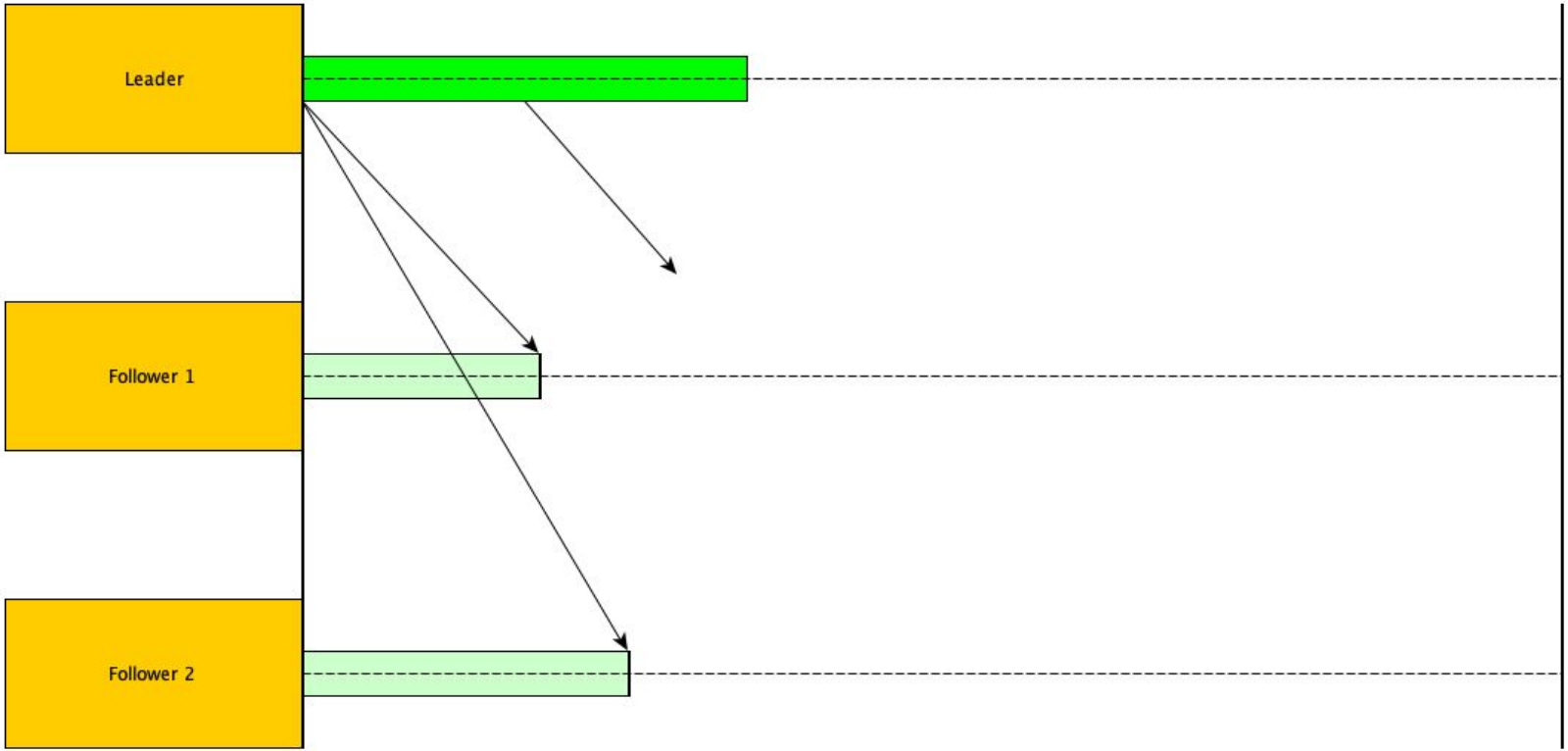
Решение

Нода, являющаяся лидером следит за обрывом соединений к/от реплик. Если количество живых соединений (живым считается двустороннее соединение) стало меньше, чем текущий кворум - лидер снимает с себя полномочия и замораживает текущие транзакции.

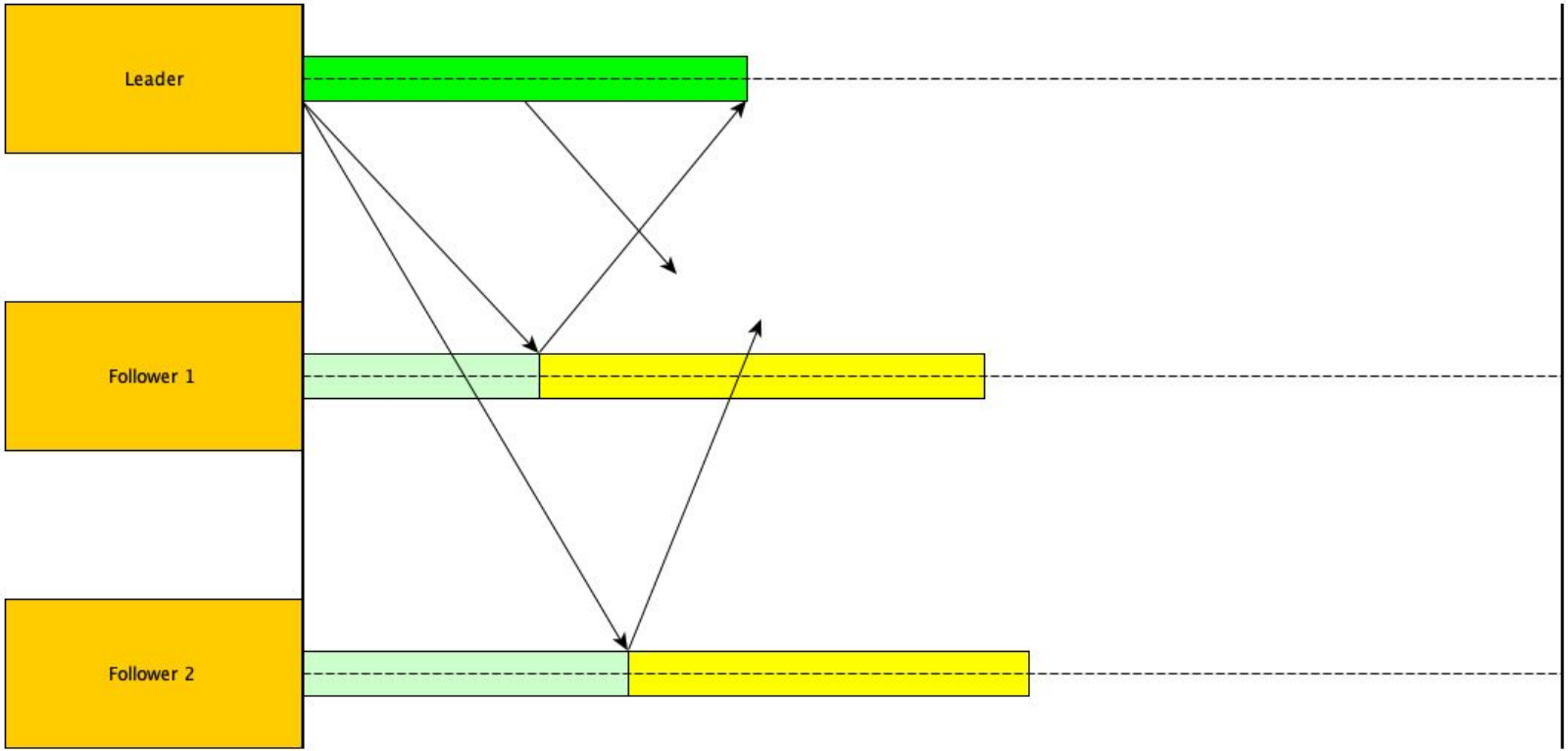
Решение



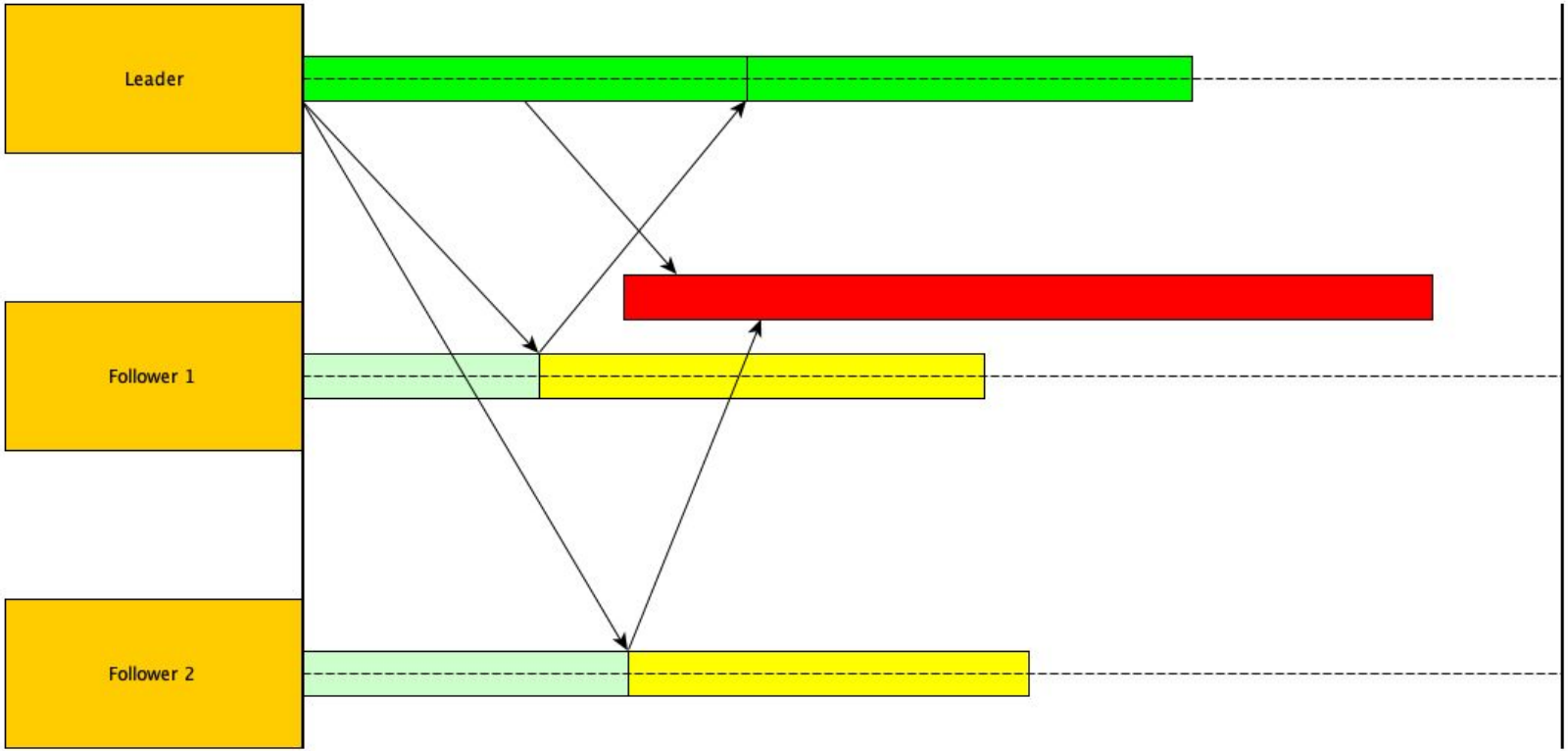
Решение



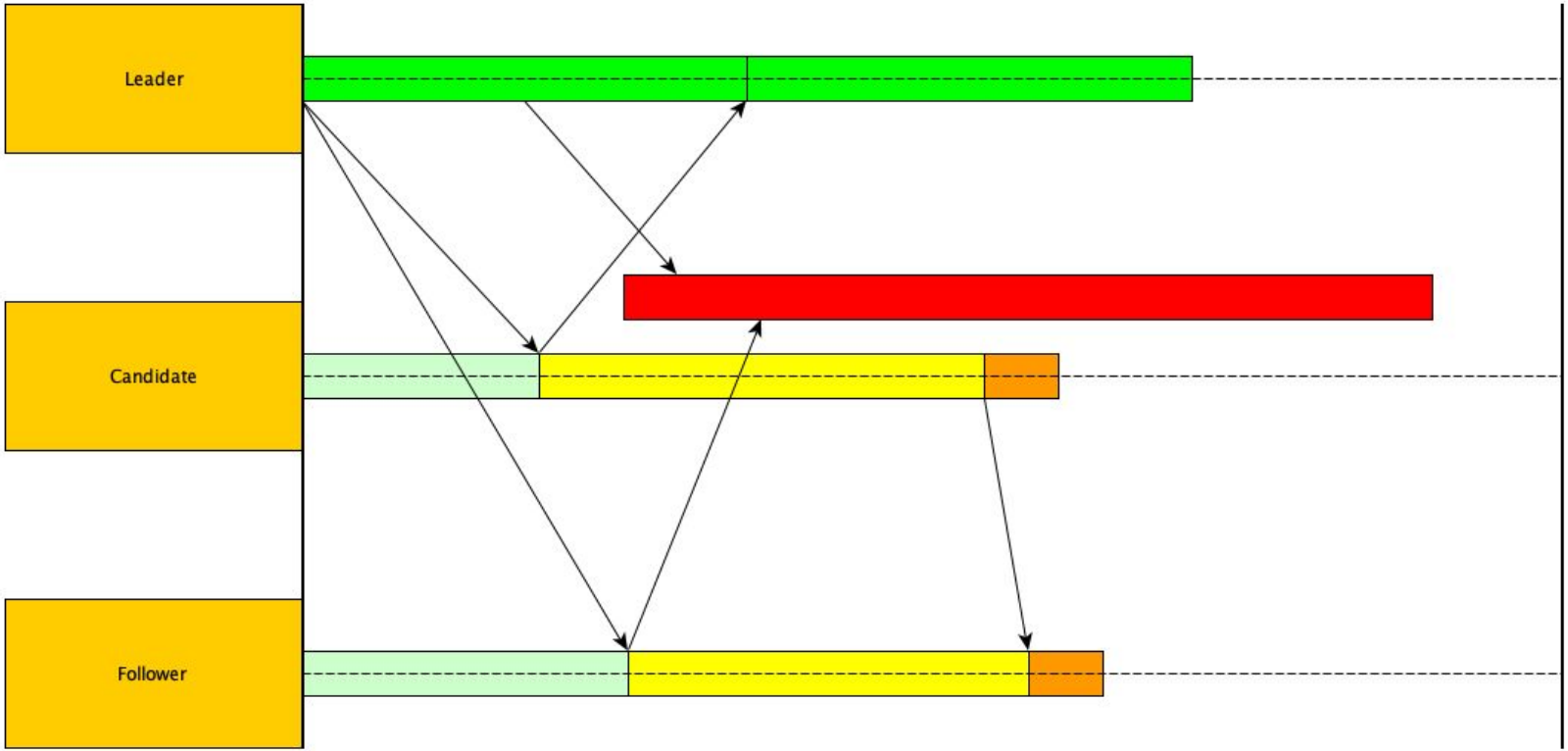
Решение



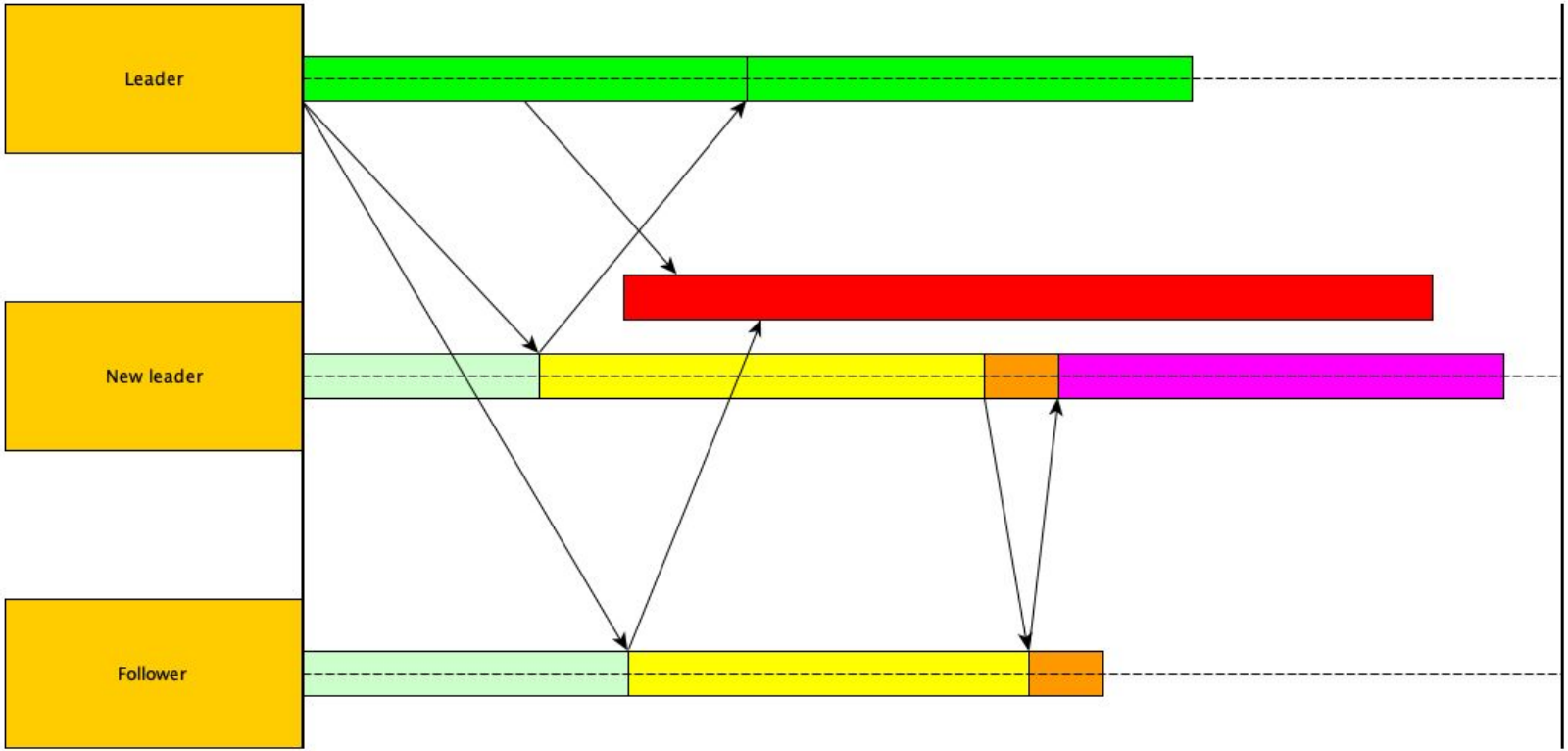
Решение



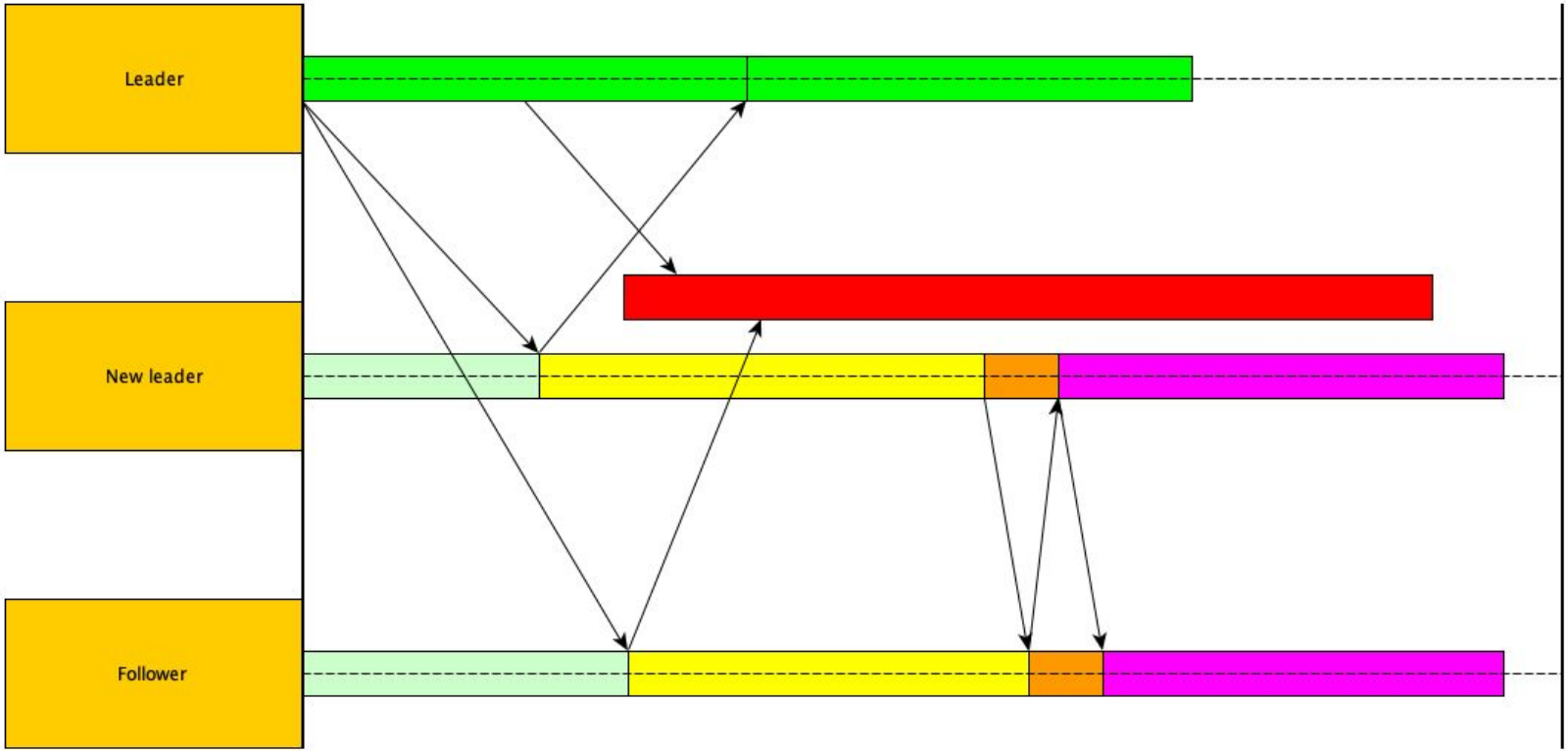
Решение



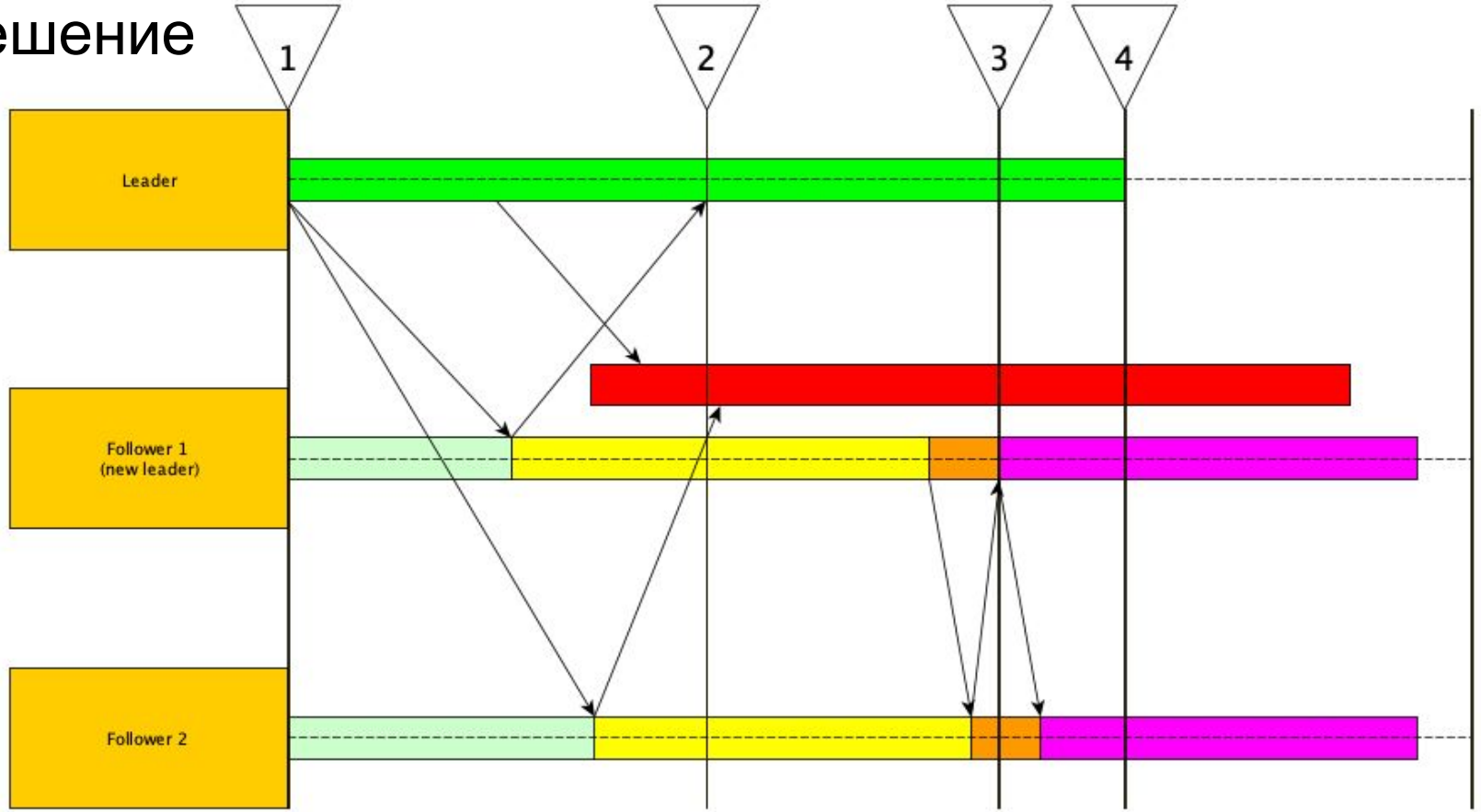
Решение



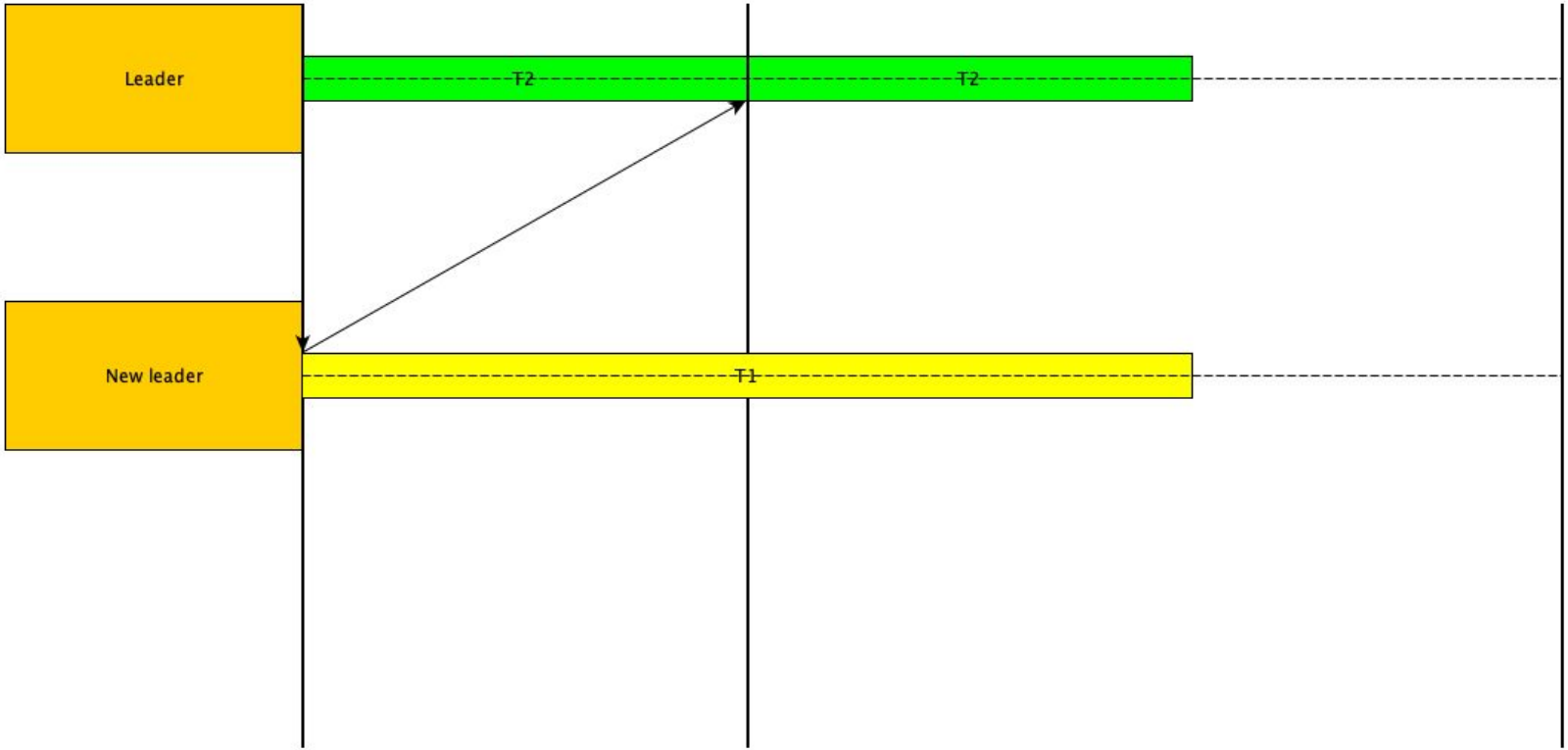
Решение



Решение



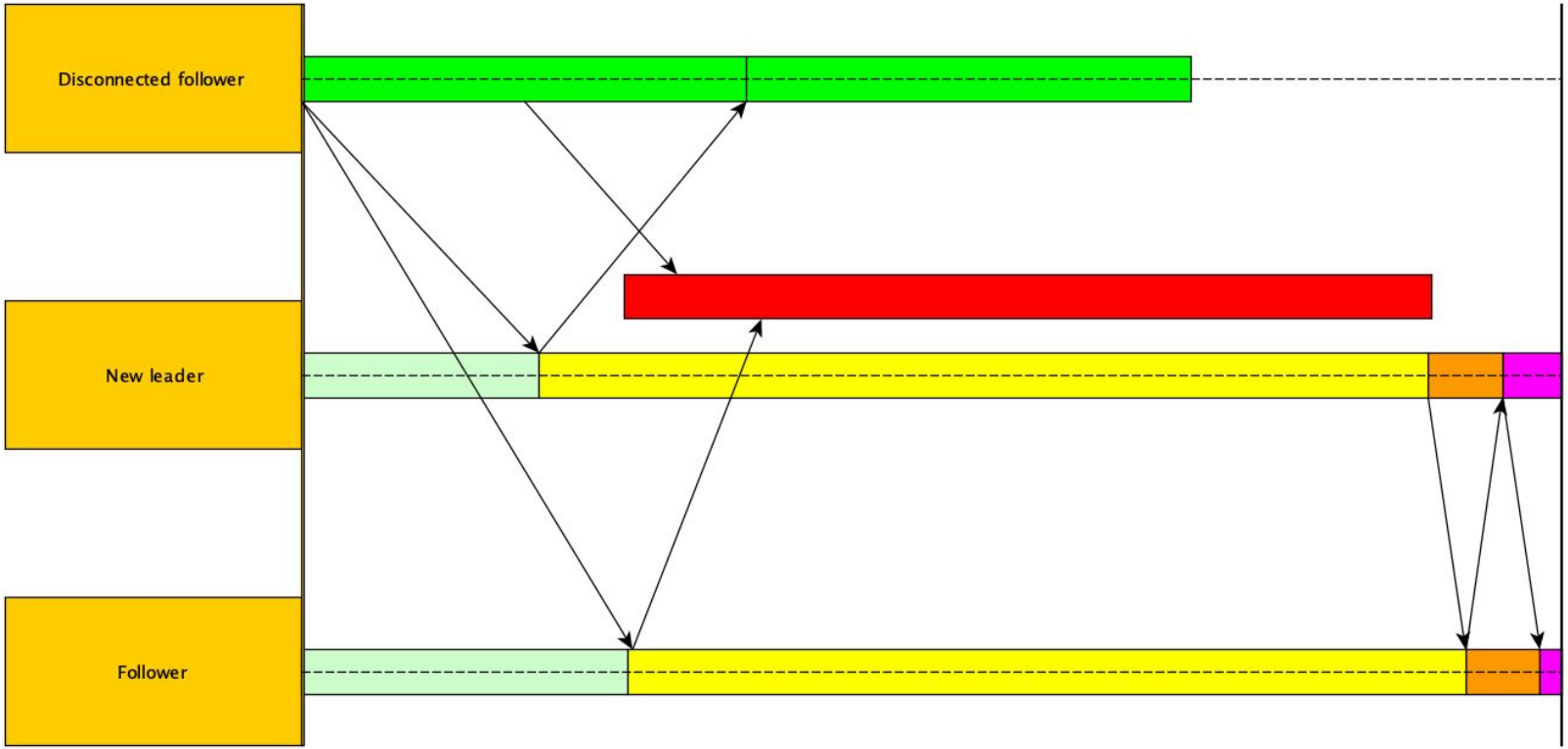
Решение



Решение

- Либо уменьшаем таймаут на лидере в два раза
- Либо увеличиваем на follower’ах в два раза

Решение



Решение

- Fencing выключен
- Слабый fencing
- Сильный fencing

У нас был план и мы его придерживались

- ✓ Raft overview
  - Journal, Term, LSN
  - Выборы
  - Гарантии
- ✓ Raft / Tarantool: особенности
- **Надстройки Raft**
  - ✓ Pre-Vote
  - ✓ Split-Vote detection
  - ✓ Fencing
  - **Manual promotion & Split-Brain detection**

Проблема

Ограниченная живучесть Raft: потеря кворума узлов приводит к недоступности кластера на запись

«Сбалансируйте 6 гвоздей так, чтобы они не касались деревянной подставки»

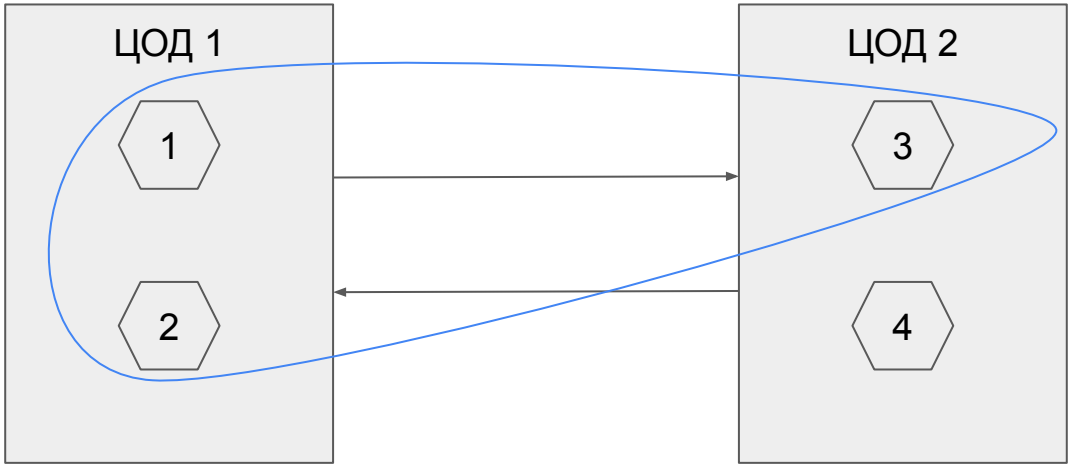
Архитектор: **Инженер:**

Проблема

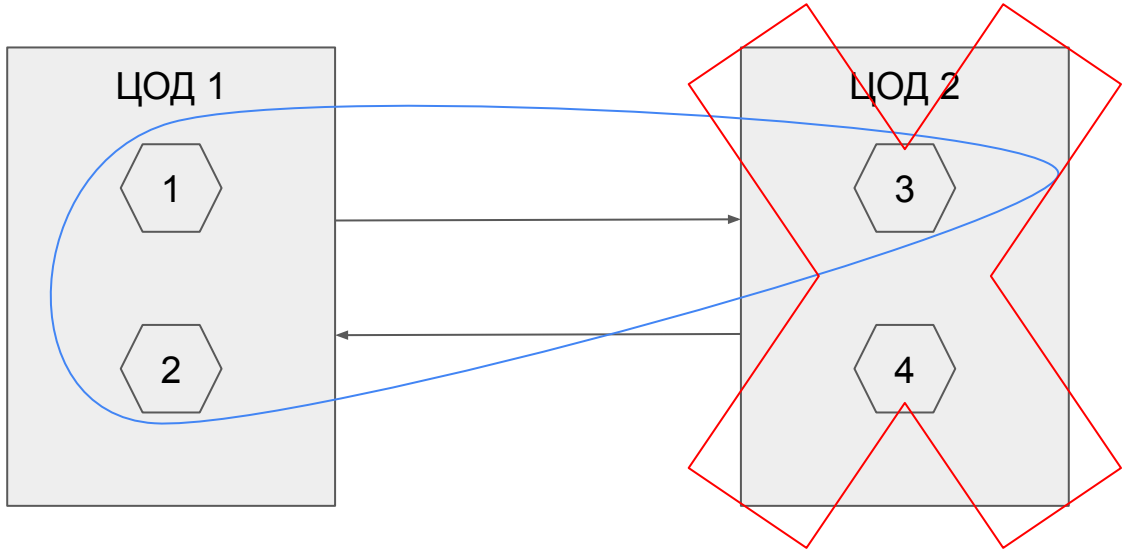
Ограниченная живучесть Raft: потеря кворума узлов приводит к недоступности кластера на запись

В экстренной ситуации может потребоваться вручную выбрать лидера, проигнорировав отсутствие кворума

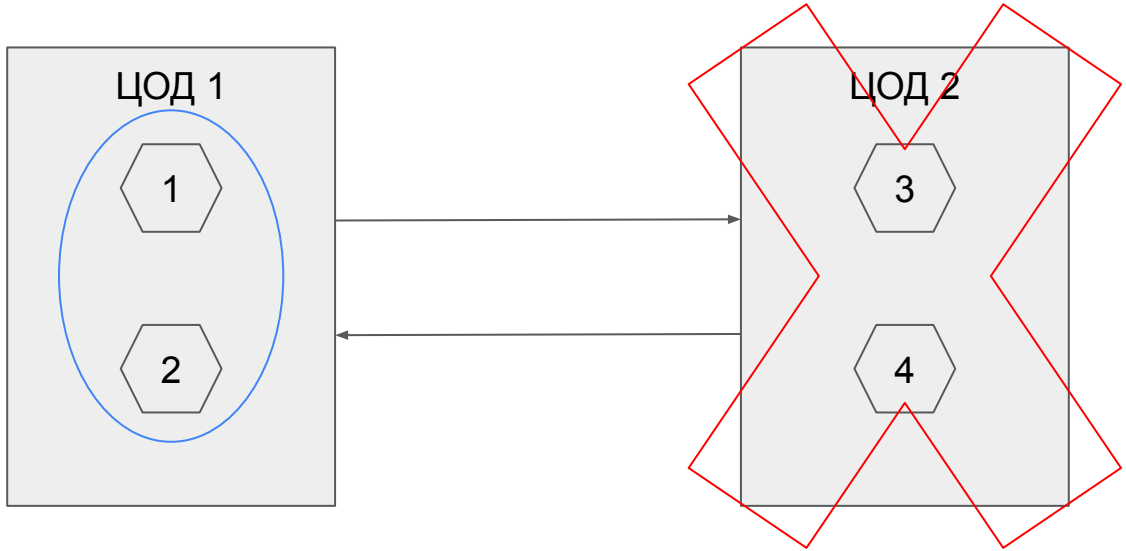
Проблема



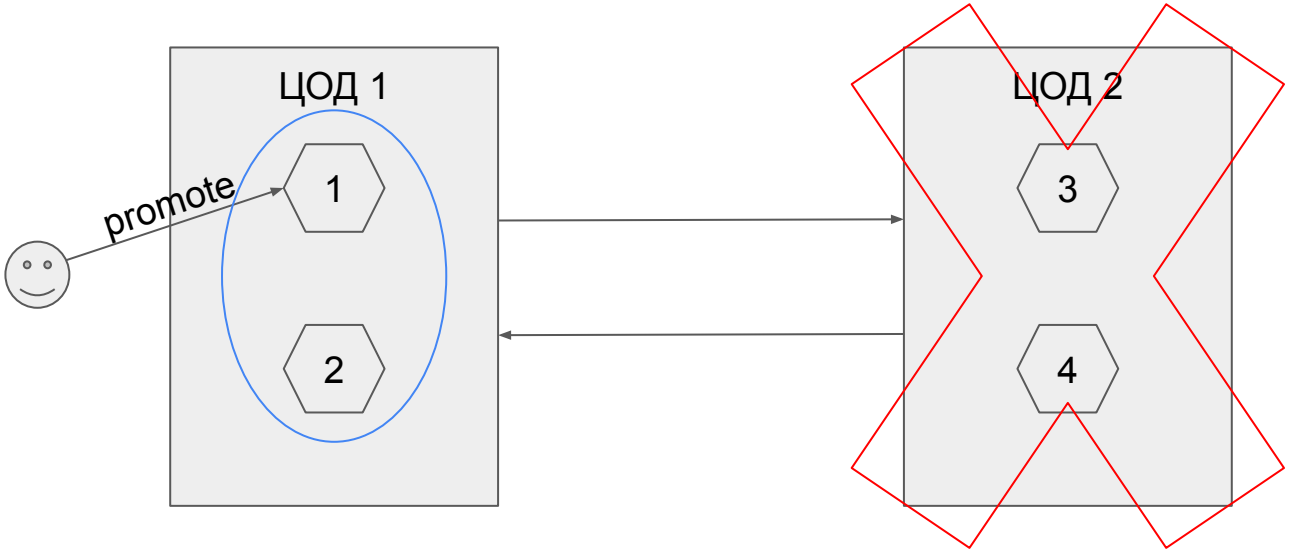
Проблема



Хочется



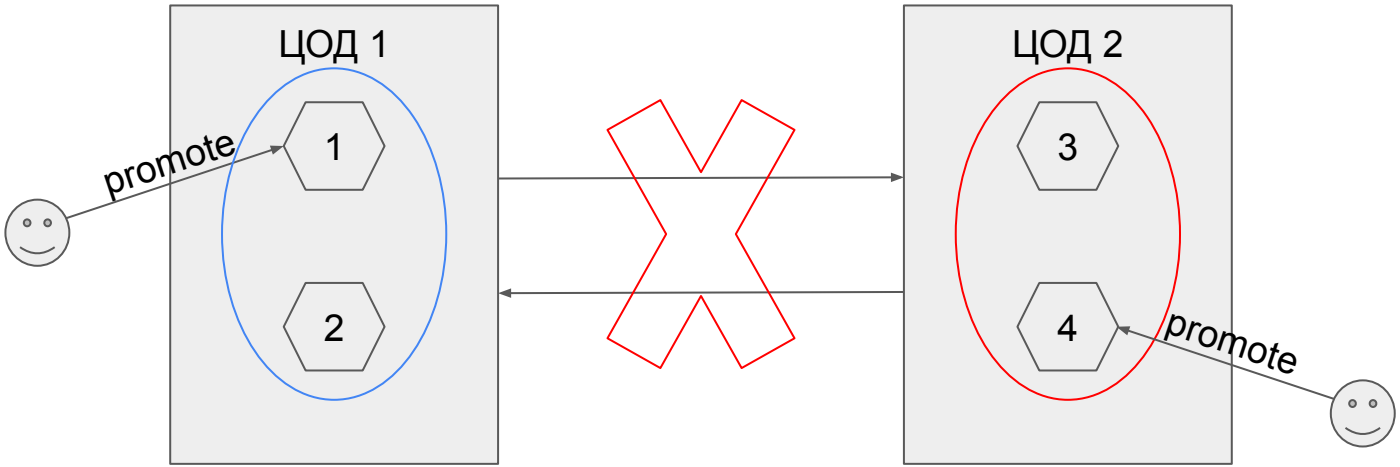
Решение



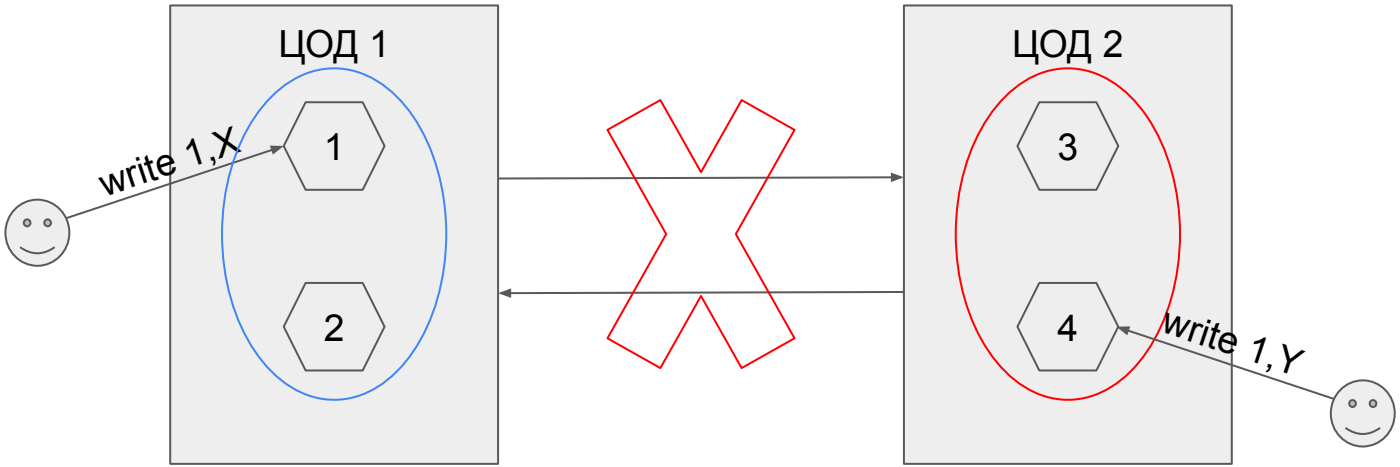
Проблема №2 (собственноручно созданная)

При ручном вмешательстве никто не застрахован от ошибок, а кластер – от сплитбрейна

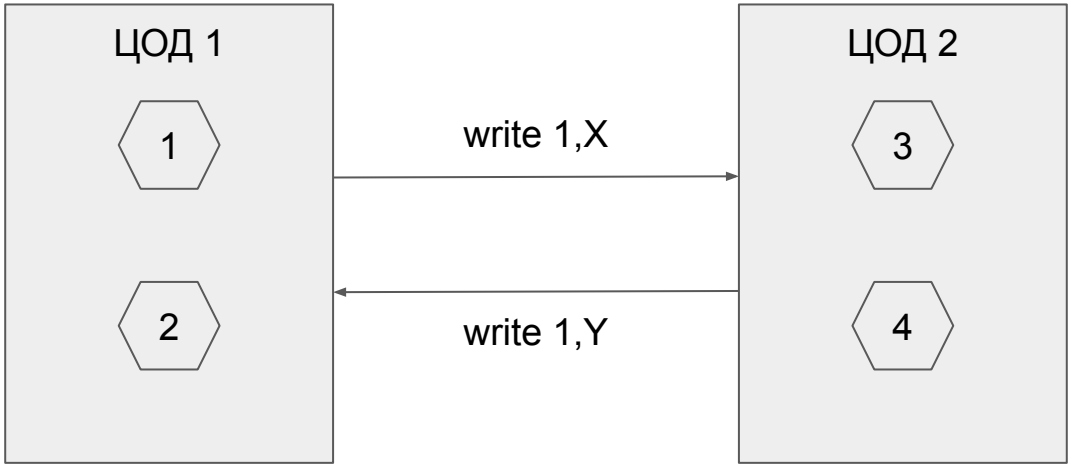
Проблема №2 (собственноручно созданная)



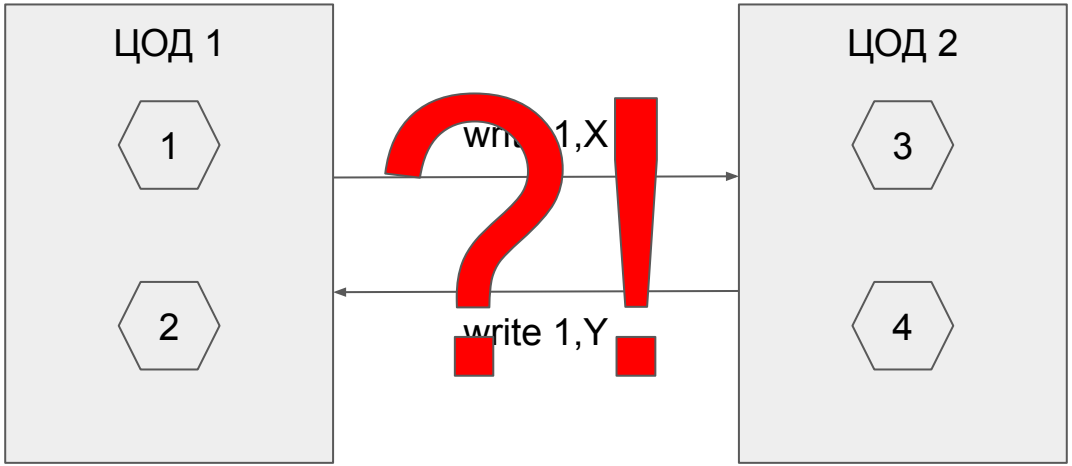
Проблема №2 (собственноручно созданная)



Проблема №2 (собственноручно созданная)



Проблема №2 (собственноручно созданная)



Как хочется

Хотим обнаруживать ошибку в момент воссоединения двух половин кластера

Как хочется

Хотим обнаруживать ошибку в момент воссоединения двух половин кластера

Обнаружив, что из-за ручного вмешательства появилось два лидера, не знавших друг о друге, не применяем никакие изменения второй половины кластера

Как хочется

Хотим обнаруживать ошибку в момент воссоединения двух половин кластера

Обнаружив, что из-за ручного вмешательства появилось два лидера, не знавших друг о друге, не применяем никакие изменения второй половины кластера

Ждём вмешательства человека, который назначит “правильную” половину. Вторая половина будет восстановлена из данных первой

Почему хотим это?

Поскольку проблема появилась из-за ручного вмешательства, решать её может только пользователь.

Почему хотим это?

Поскольку проблема появилась из-за ручного вмешательства, решать её может только пользователь.

Система не имеет права вмешиваться

Почему хотим это?

Поскольку проблема появилась из-за ручного вмешательства, решать её может только пользователь.

Система не имеет права вмешиваться

Система должна сохранить обе версии данных

Решение

Мультиплексированный журнал – после воссоединения кластера получаем недостающие “кусочки” журнала от второго лидера

Решение

Мультиплексированный журнал – после воссоединения кластера получаем недостающие “кусочки” журнала от второго лидера

Данные, применённые любым лидером предваряются специальной записью PROMOTE, указывающей на этого лидера и терм

Решение

Мультиплексированный журнал – после воссоединения кластера получаем недостающие “кусочки” журнала от второго лидера

Данные, применённые любым лидером предваряются специальной записью PROMOTE, указывающей на этого лидера и терм

Первый же PROMOTE, нарушающий линейную историю, свидетельствует о том, что был сплит-брейн

Решение

Мультиплексированный журнал – после воссоединения кластера получаем недостающие “кусочки” журнала от второго лидера

Данные, применённые любым лидером предваряются специальной записью PROMOTE, указывающей на этого лидера и терм

Первый же PROMOTE, нарушающий линейную историю, свидетельствует о том, что был сплит-брейн

Обнаружив сплит-брейн, разрываем соединение, не применяя последующие изменения

У нас был план и мы его придерживались

- ✓ Raft overview
  - Journal, Term, LSN
  - Выборы
  - Гарантии
- ✓ Raft / Tarantool: особенности
- ✓ Надстройки Raft
  - Pre-Vote
  - Split-Vote detection
  - Fencing
  - Manual promotion & Split-Brain detection

Ожидали от Raft

- Один лидер на терм == один лидер в данный момент времени
- Лидер постоянен
- Быстрая смена лидера

Получили

- Один лидер на терм == один лидер в данный момент времени – Fencing
- Лидер постоянен – Pre-Vote, Fencing
- Ускорение смены лидера – Split-Vote detection
- Защита от последствий ручного вмешательства – Split-brain detection

Сергей Петренко

Tarantool

s.petrenko@corp.mail.ru

tg: sergepetrenko

Борис Степаненко

Tarantool

b.stepanenko@vk.team

tg: Grafin1992

Tarantool Hydra

Спасибо за внимание!

https://github.com/tarantool/tarantool