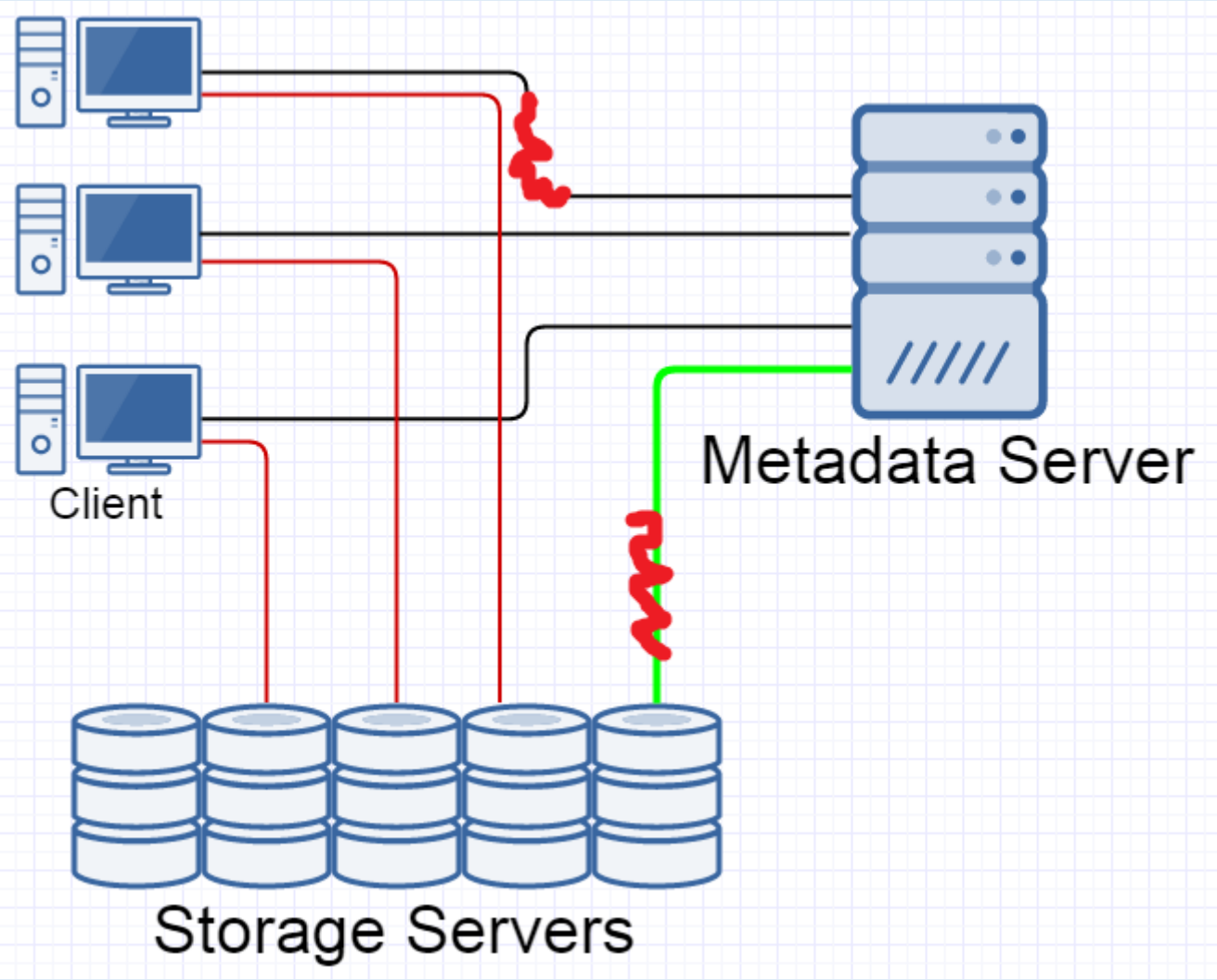
Основы построения файловых систем

Консенсус в распределённой системе

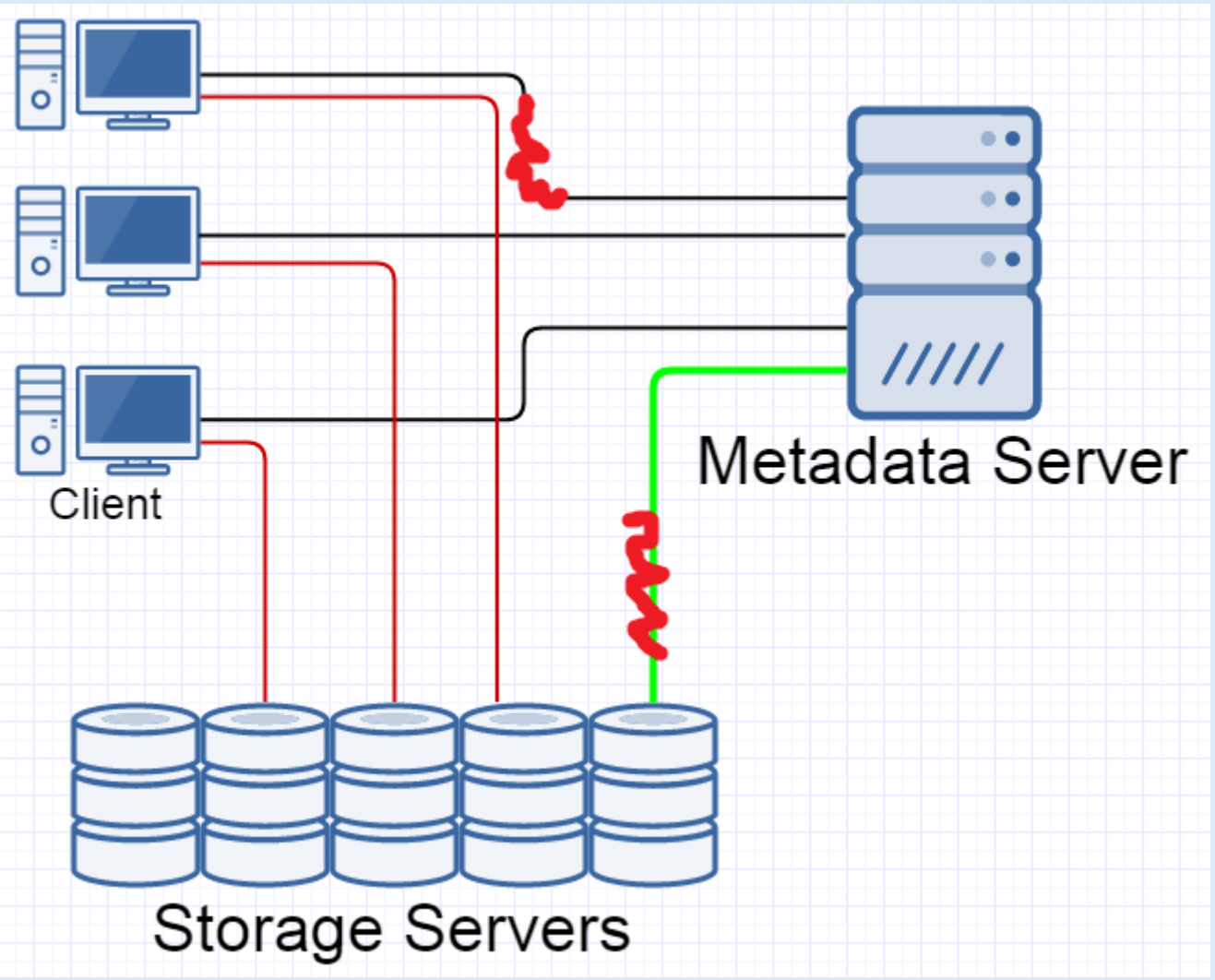
Сегодня мы поговорим о том, как сделать надёжный распределённый конечный автомат.

Metadata server in Parallel NFS



Проблема: metadata server, даже если справляется с нагрузкой, является «single point of failure», т.е. при отказе одного MDS становится недоступной вся система.

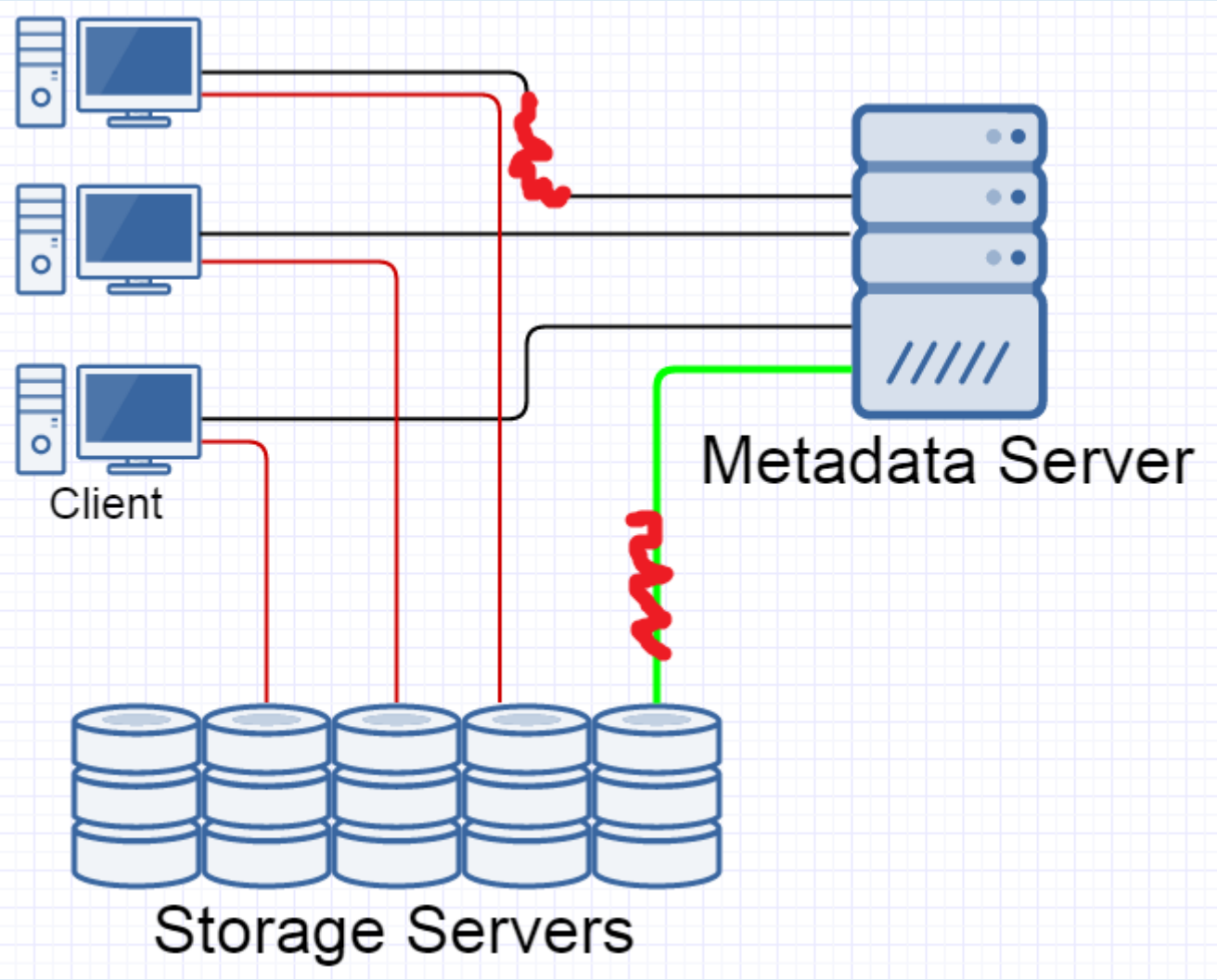
Metadata server in Parallel NFS



Проблема: metadata server, даже если справляется с нагрузкой, является «single point of failure», т.е. при отказе одного MDS становится недоступной вся система.

Идея: давайте сделаем много экземпляров MDS.

Metadata server in Parallel NFS



Проблема: metadata server, даже если справляется с нагрузкой, является «single point of failure», т.е. при отказе одного MDS становится недоступной вся система.

Идея: давайте сделаем много экземпляров MDS.

Проблема: как обеспечить согласование состояния на разных MDS?

Простой пример распределённой системы: SMP

Классический пример несогласованных картин мира:

Thread 0	Thread 1
++x;	--x;

Простой пример распределённой системы: SMP

Классический пример несогласованных картин мира:

Thread 0	Thread 1
++x;	--x;

Сценарий 0		Сценарий 1	
CPU0	CPU1	CPU0	CPU1
R0 <- x		R0 <- x	
	R1 <- x	Inc R0	
Inc R0	Dec R0	R0 -> x	
R0 -> x			R0 <- x
	R0 -> x		Dec R0
			R0 -> x

Свой вклад ещё вносят: out-of-order execution, write reordering, write combining, CPU caches.

Простой пример распределённой системы: SMP

Проблемы CPU:

- Одновременный доступ,
- Переупорядочивание,
- Актуальная версия данных может располагаться только в кеше одного из процессоров.

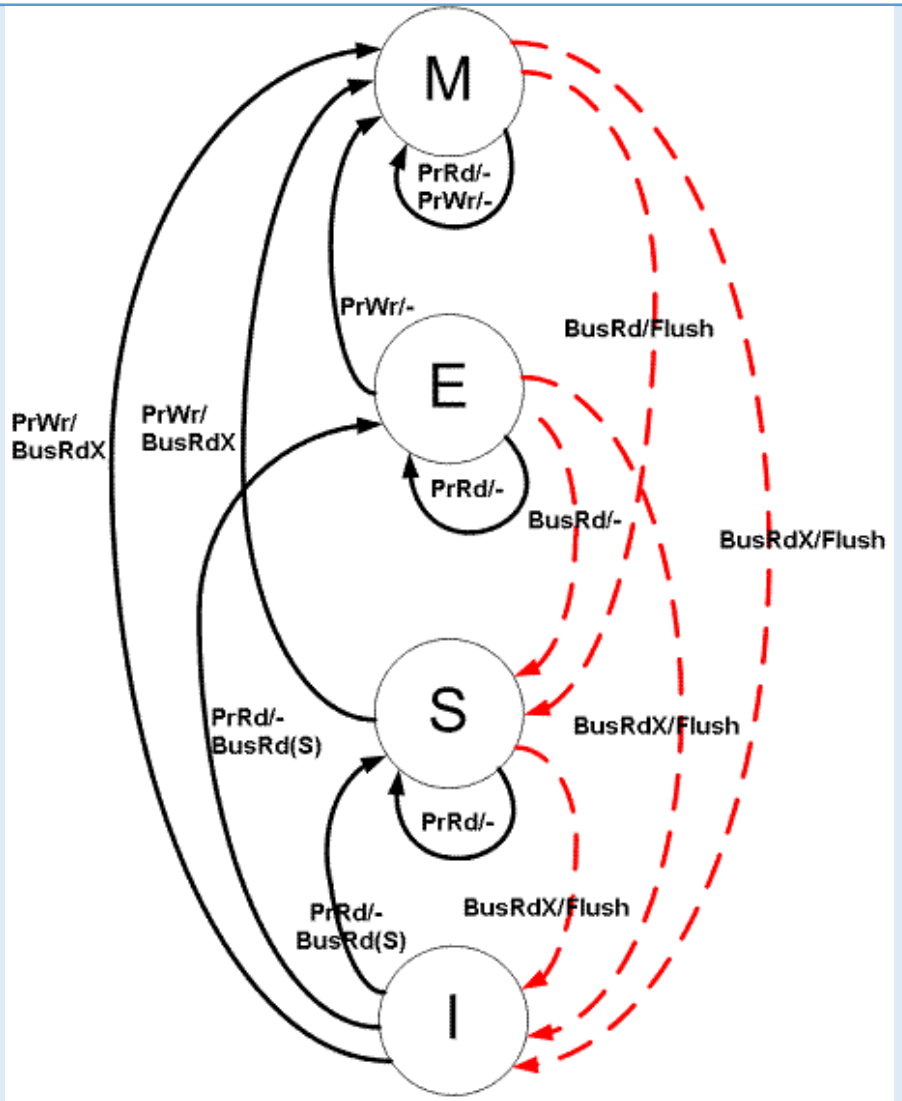
Простой пример распределённой системы: SMP

Проблемы CPU:	
• Одновременный доступ,	Atomic operations, transactional memory (более простой вариант – load-linked / store-conditional)
• Переупорядочивание,	Memory barriers
• Актуальная версия данных может располагаться только в кеше одного из процессоров.	Cache coherency protocol

Простой пример распределённой системы: SMP

Modified	Актуальные данные только в данном кеше
Exclusive	Область памяти закеширована только в этом кеше, но значение в кеше совпадает со значением в памяти
Shared	Область памяти может быть закеширована несколькими кешами, но значения в кешах совпадают со значением в памяти
Invalid	Линейка в кеше пуста

Протокол полагается на возможность подслушивать запросы **всех** процессоров.



https://en.wikipedia.org/wiki/MESI_protocol

Консенсус в системе с ненадёжными участниками

Теперь рассмотрим систему, узлы в которой могут на время становиться недоступными, и сообщения в которой могут теряться.

Чтобы реализовать распределённый FSM, хватит уметь распределённо выбирать одно из предложений, которые делают участники:

- Предложения имеют вид «шаг FSM номер N – это переход из состояния X в состояние Y»,
- На добавление/удаление участников тоже можно смотреть на шаг FSM.

Роли в PAXOS

- Proposer
- Acceptor
- Learner

Предлагаемые значения имеют вид (n, v), где n – натуральное число (время на proposer’e), v – значение, которое собственно предлагается.

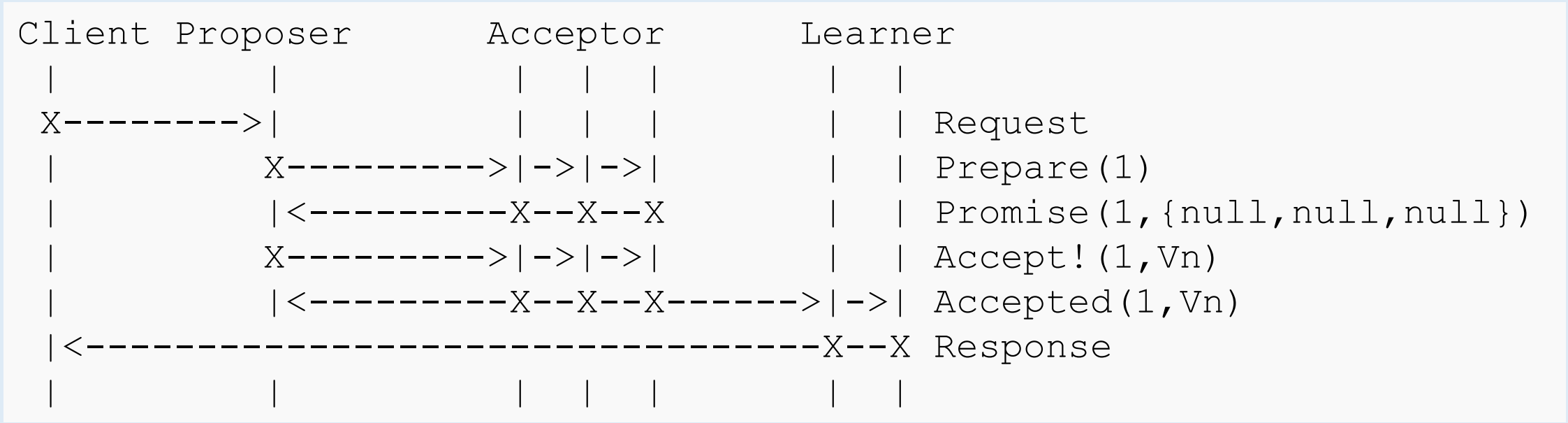
Для достижения консенсуса необходимо подтверждение принятия от большинства участников.
Замечание: большинством можно называть любое множество из семейства подмножеств acceptor’ов с попарно непустыми пересечениями.

Роли в PAXOS

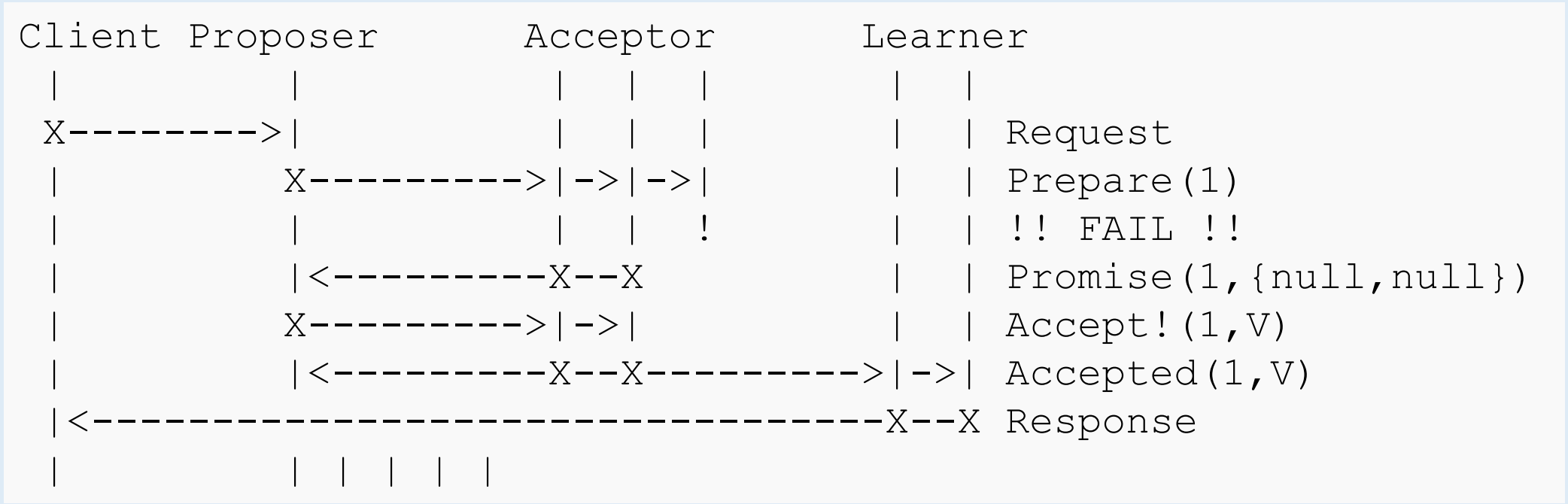
Proposer	Acceptor
Выбирает новое значение **n** и рассылает большинству acceptor'ов запрос prepare(n)	Если полученное в запросе prepare(n) значение больше, чем максимальное ранее полученное, то отвечает proposer'у, обещая не принимать значения с меньшими **n**, а также отсылает принятое им ранее значение **v** (или null).
Получив ответ от большинства, рассылает большинству запрос accept(n, v), где **v** – значение с наибольшим номером, полученное от acceptor'ов; если же acceptor'ы ещё не приняли никакого значения, то **v** – это значение, которое хотел предложить proposer.	Получив запрос accept (n, v), принимает значение v, если не получал запросов prepare(m) с большими номерами. Сообщает об этом learner'ам, или некоторым выделенным learner'ам.

- http://research.microsoft.com/en-us/um/people/lamport/pubs/lamport-paxos.pdf
- http://research.microsoft.com/en-us/um/people/lamport/pubs/paxos-simple.pdf

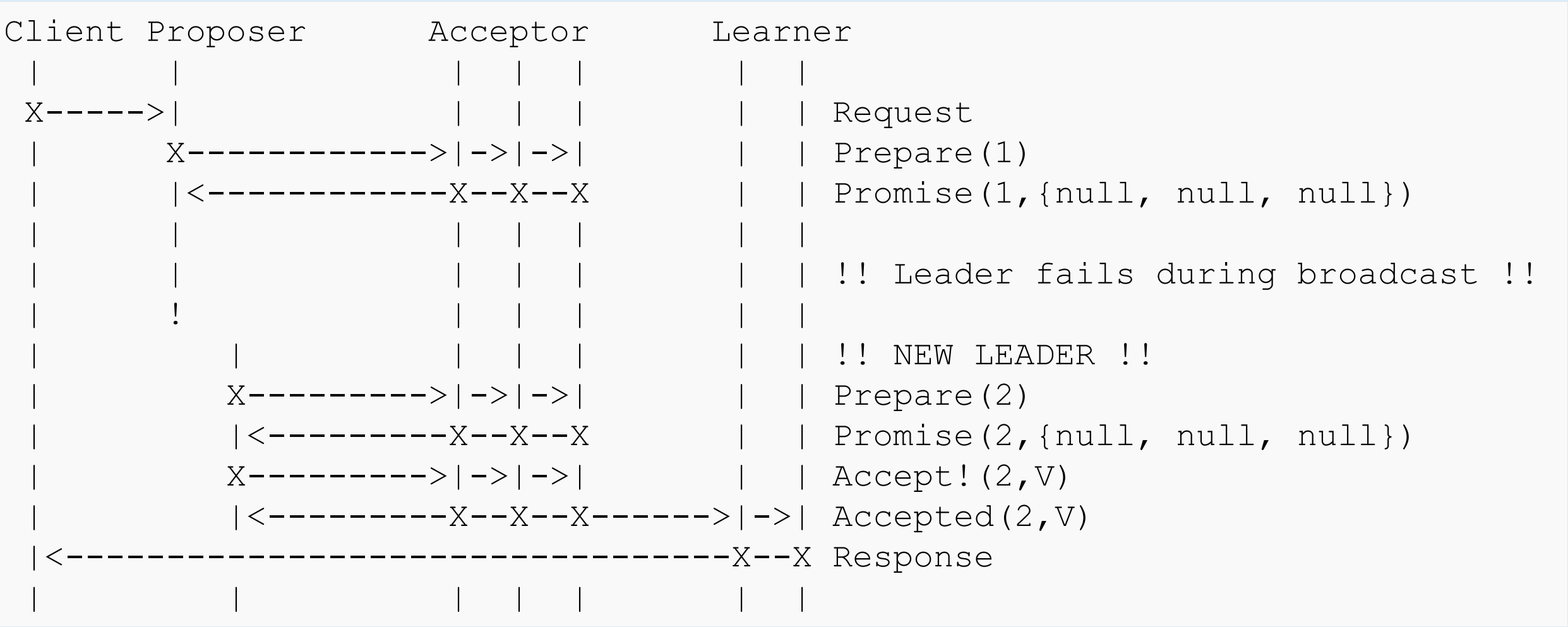
PAXOS, примеры (normal workflow)



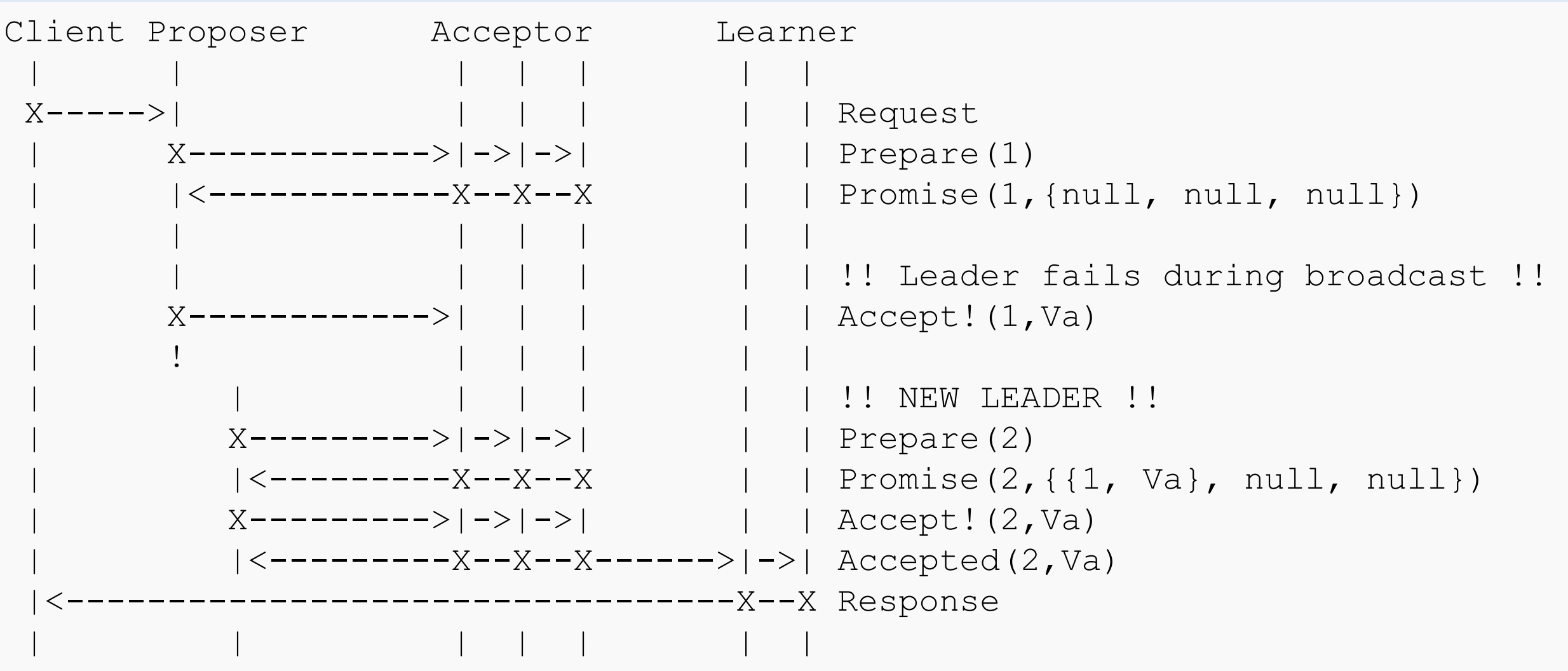
PAXOS, примеры (failure of an acceptor)



PAXOS, примеры (failure of a proposer)



PAXOS, примеры (failure of a proposer)



PAXOS

- Прогресс гарантируется только при наличии выделенного proposer’а.
- У каждого участника должен быть журнал.
- Снимки состояния участников.

CAP-теорема

Желаемые свойства распределённой системы:

- C – Consistency,
- A – Availability,
- P – Partition tolerance.

В грубом приближении, CAP-теорема – это утверждение о том, что все три свойства одновременно обеспечить не получится.

Примеры:

- Базы данных – C & P,
- Amazon S3 – A & P.