Redis Cluster

Come funziona, come fallisce.

Cos’e’ la performance?

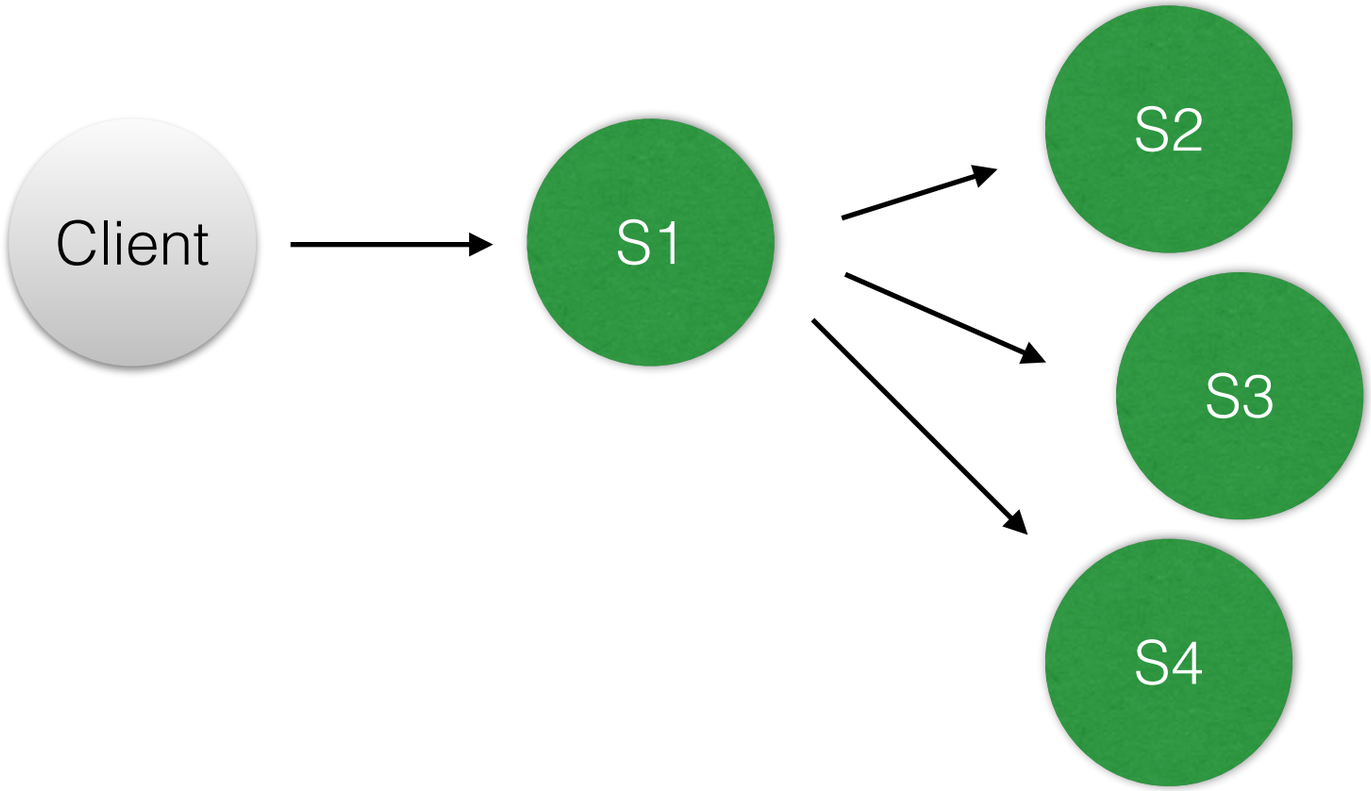
- Bassa latenza.
- IOPS.
- Qualita’ delle operazioni.
- Redis su singolo nodo e’ performante…

Go Cluster

- Redis Cluster non poteva tradire Redis.
- Anche se bisogna scendere a compromessi.
- CAP? Fare il merge dei valori? Mirare alla strong consistency? Come replicare i dati?

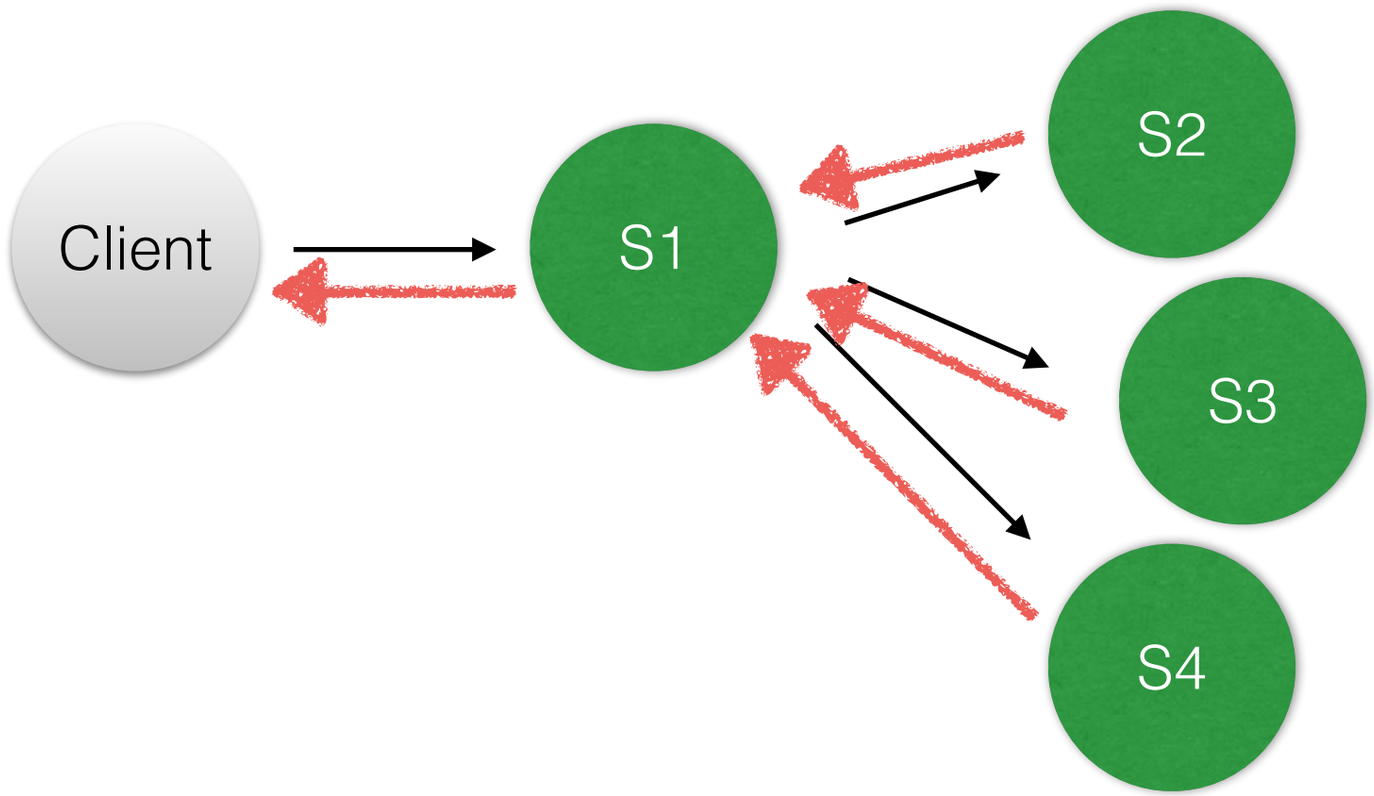
I sistemi CP

CAP: paghi la consistenza con le performance.

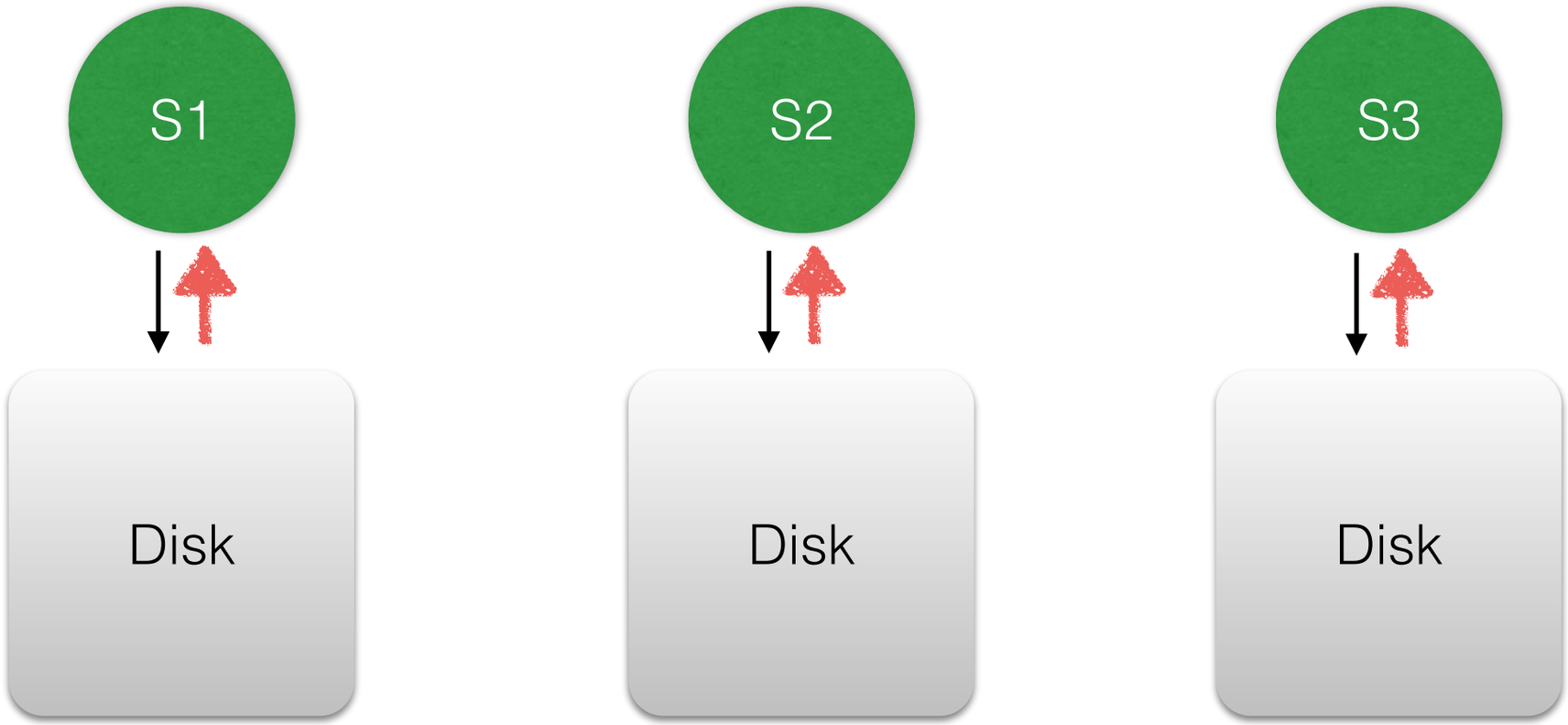


I sistemi CP

Risposta dopo ACK della maggioranza.



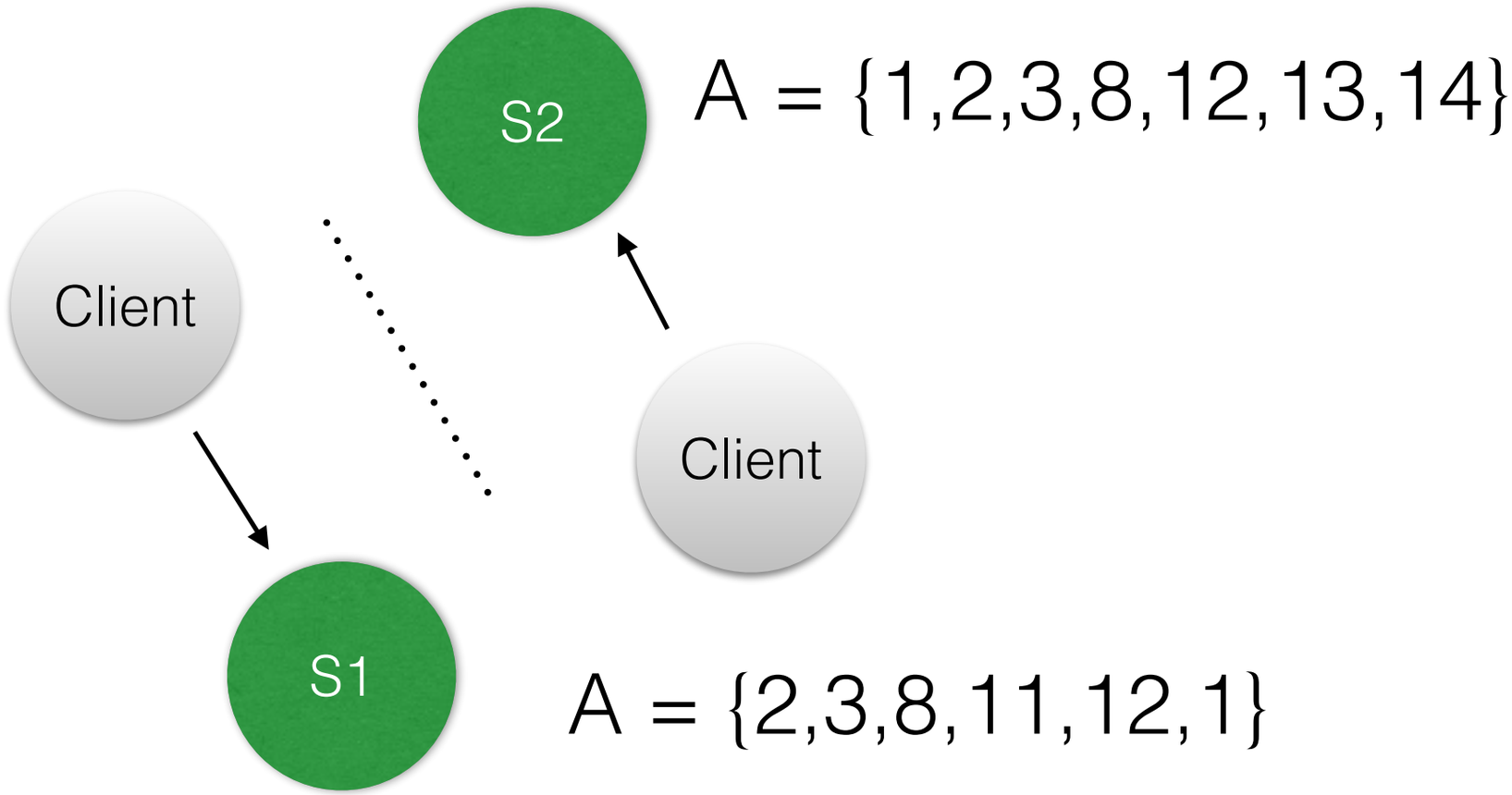
Non e’ cosi’ semplice.



Il disco di ogni nodo fa parte del nostro sistema distribuito in ogni Modello di Sistema sano.

I sistemi AP

Consistenza eventuale = Merging.



Gli altri CP…

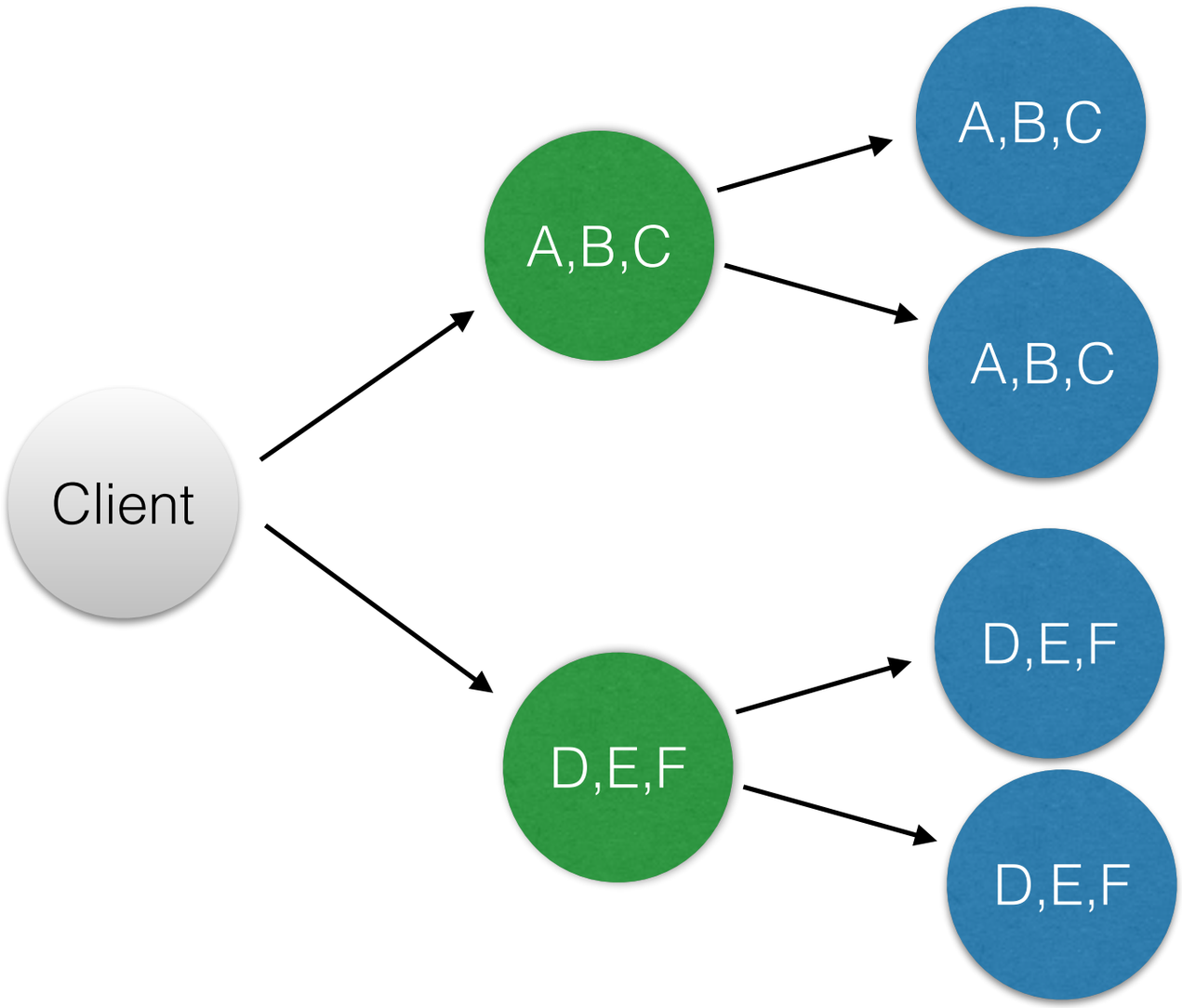
- La “C” di CAP e’ una consistenza “stretta”.

- Ma non e’ l’unica.

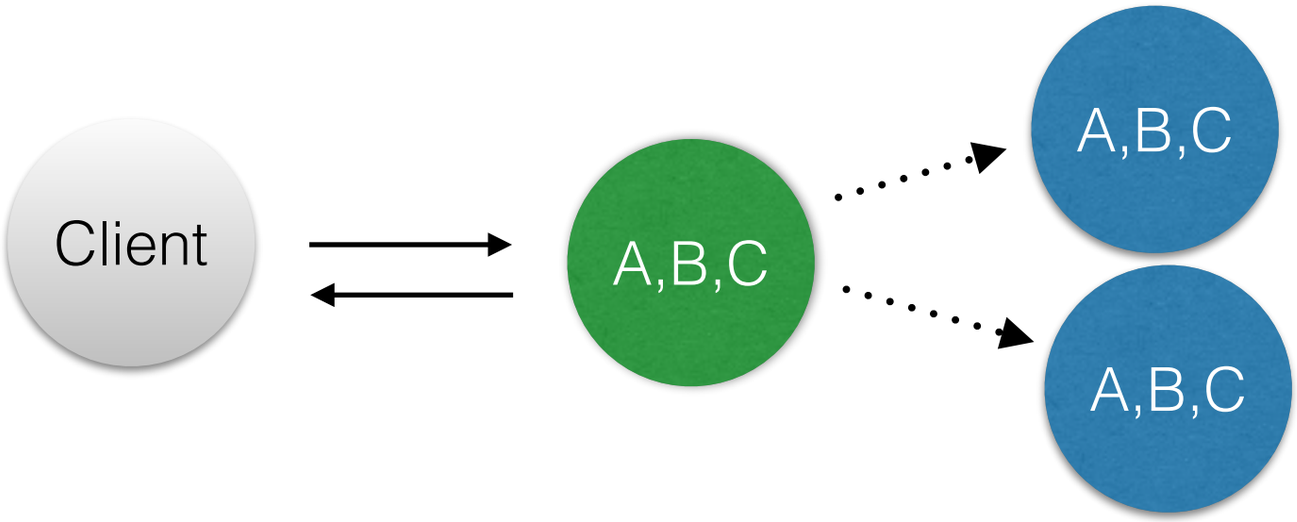
- La consistenza e’ in realta’ il contratto tra il database e il client…

Redis Cluster

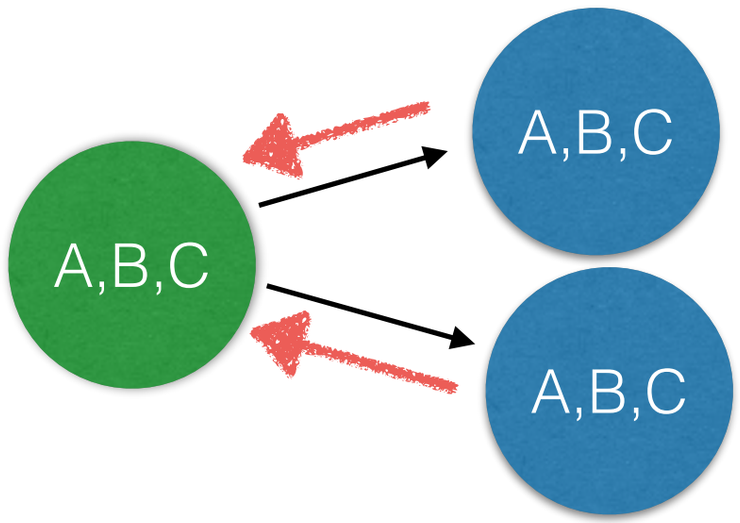
Partizionamento e Replicazione (asincrona).



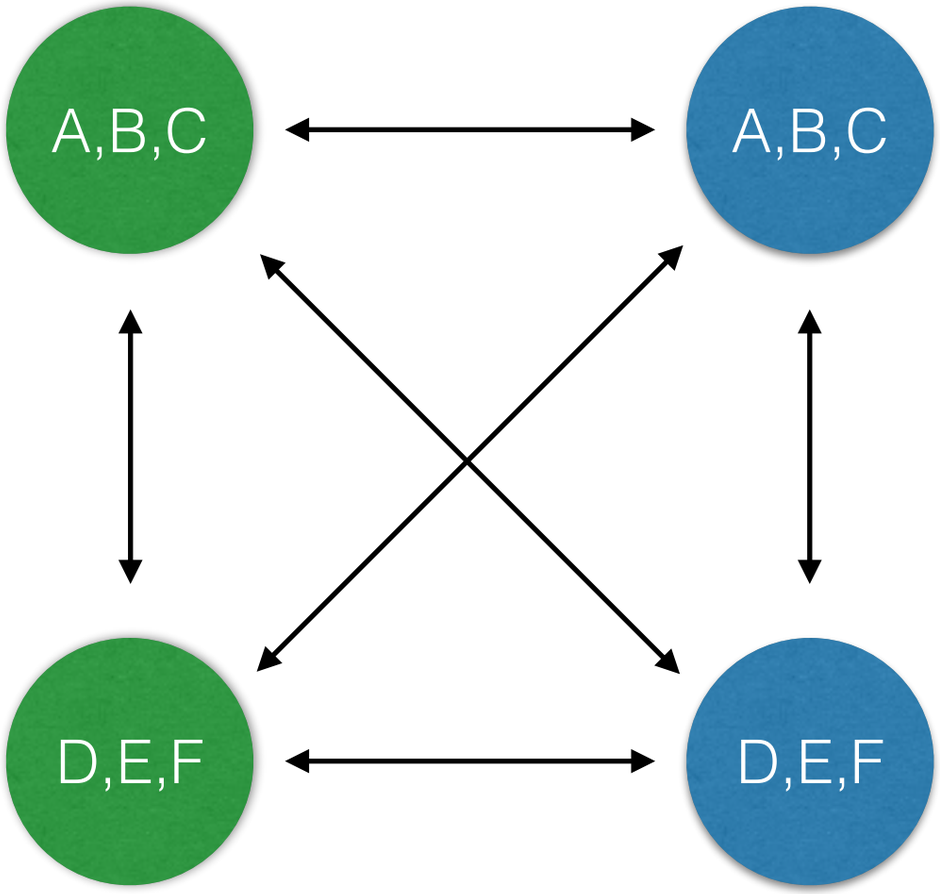
Replicazione asincrona



ACK asincrono

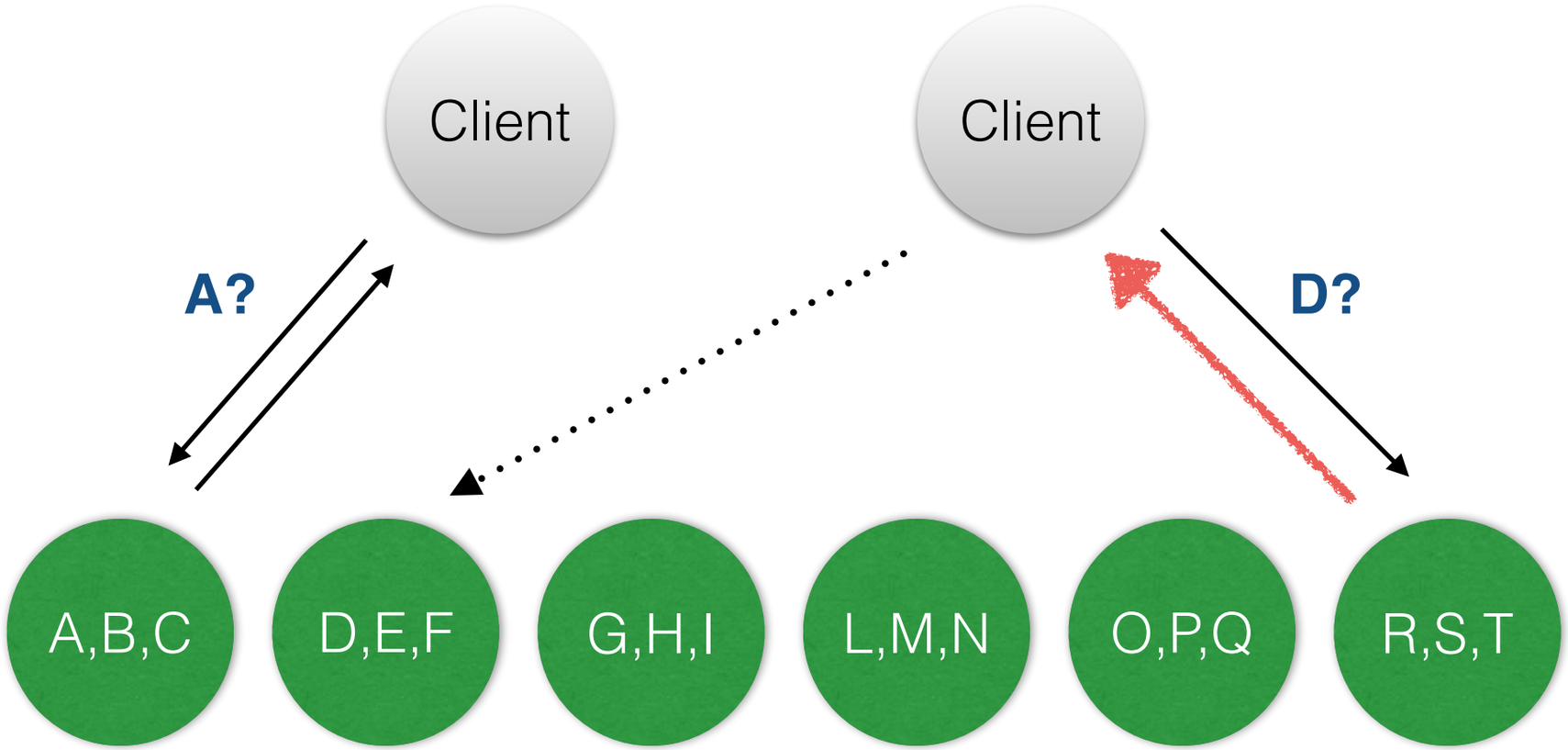


Full Mesh



- Heartbeats.
- Gossip.
- Consenso per failover.
- Propagazione configurazione.

Proxyless + Redirezioni



Failure detection

- Utilizza il gossip per arrivare ad un consenso “informale”.
- Trigger per il failover (che invece usa un consenso forte).
- Stati fondamentali: PFAIL -> FAIL

Gestire i metadata

- Dopo il fallimento, segue il failover.
- Il cluster necessita di una visione coerente.
- Chi serve questo slot ora?
- Cosa accade durante le partizioni?

Raft e il failover

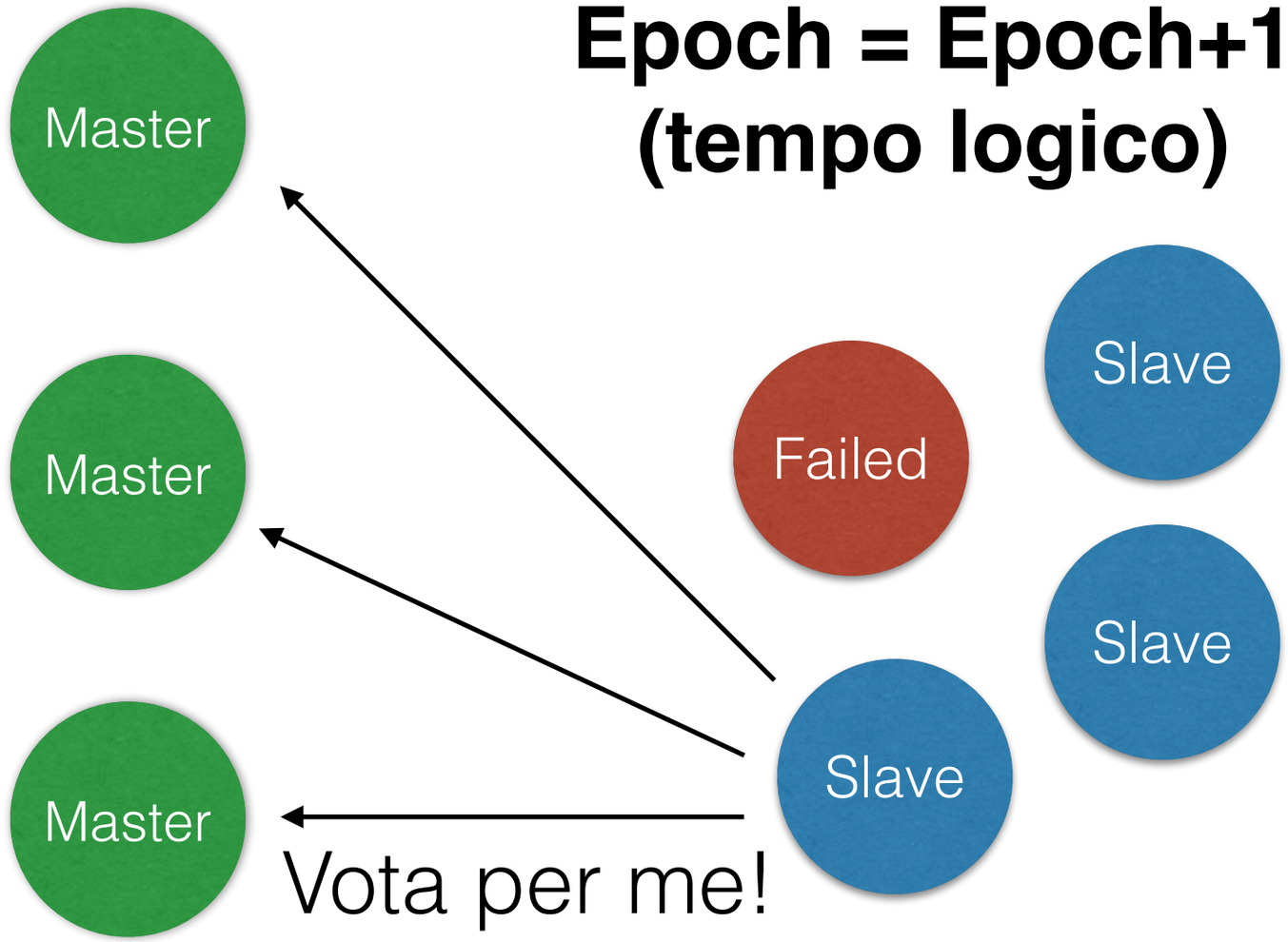
- Questi problemi sono risolti usando un “pezzo” di **Raft**.

- Raft e’ un algoritmo distribuito del consenso, come Paxos, ma fatto di parti separate e chiare.

- Il paper originale di Raft e’ gia’ una pietra miliare.

- Ma tutto Raft e’ troppo per Redis Cluster…

Failover: slave vincente



Sembra facile…

- Perche’ non abbiamo bisogno di usare Raft completo?

- Siamo forse raccomandati?

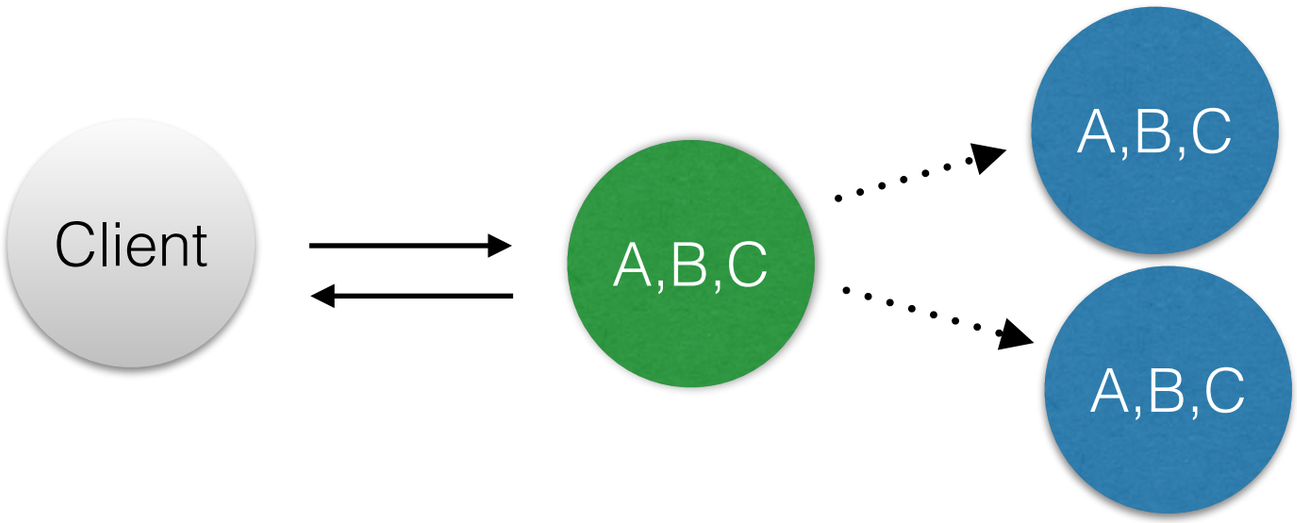
- L’essenza e’, possiamo rimpiazzare tutto lo “stato” per uno slot con un solo messaggio.

- Vedremo se funziona cosi’ bene con Sentinel :-)

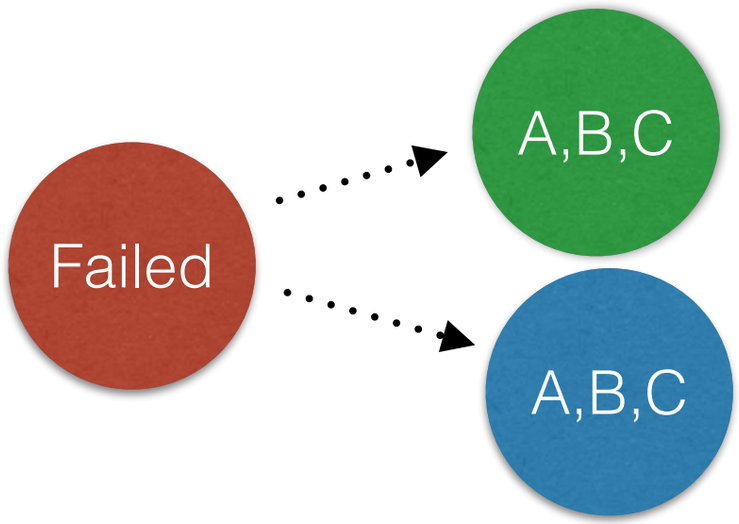
Propagazione

- Dopo il failover la configurazione viene spedita a tutti i nodi.
- Se ci sono partizioni non importa perche' viene rispedita per sempre a tutti i nodi non aggiornati.
- Quella con epoch maggiore vince sempre.

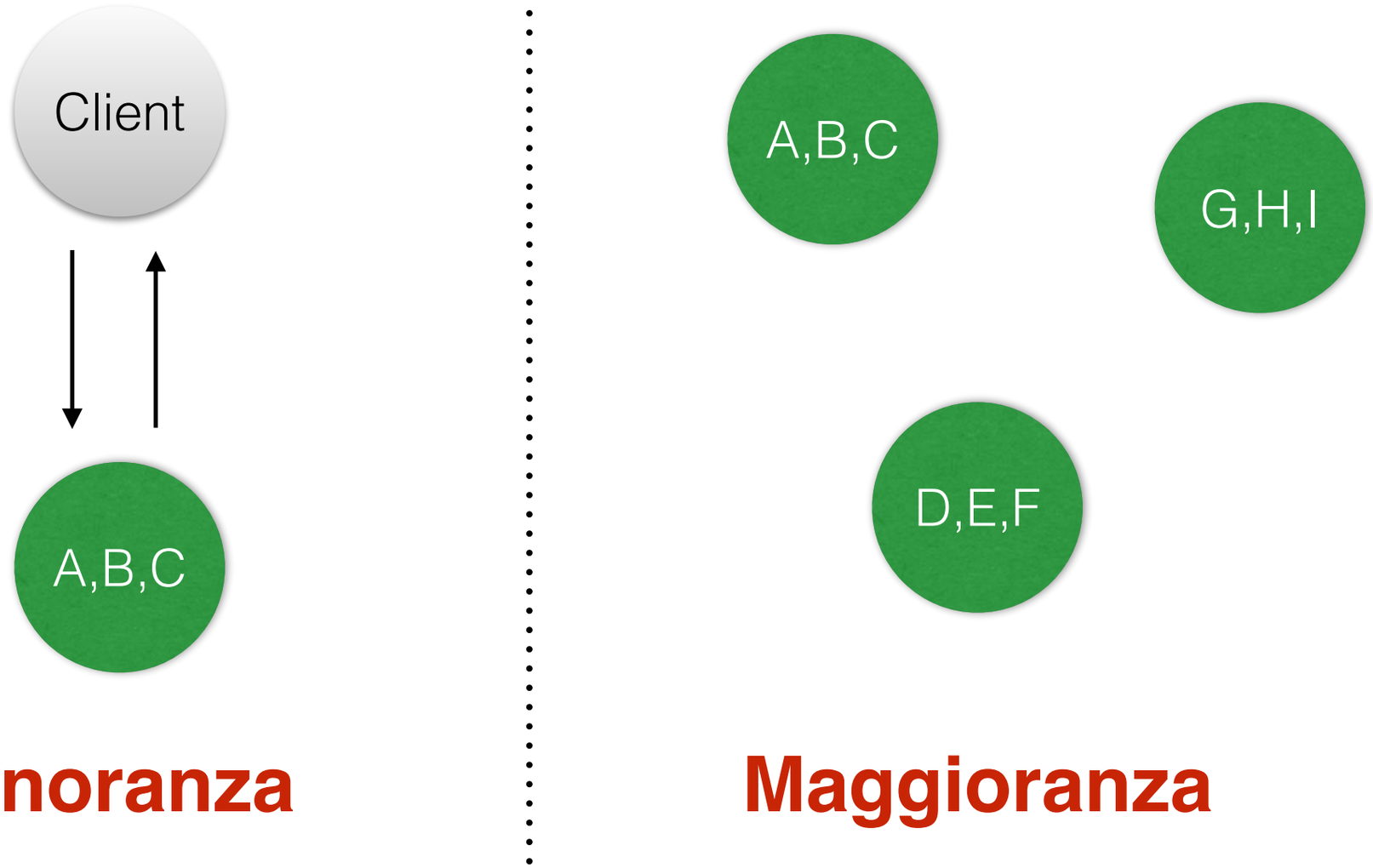
Come fallisce… #1



Scrittura persa…



Come fallisce… #2



Ask me anything

- In realta' ci sono un sacco di altri dettagli…
- http://redis.io/topics/cluster-spec
- Domande?