state-machine replication for planet-scale systems

Vitor Enes, Carlos Baquero, Tuanir França Rezende,
Alexey Gotsman, Matthieu Perrin, Pierre Sutra

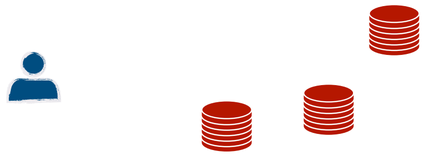
29 Apr. 2020 @ EuroSys'20

state-machine replication for planet-scale systems

state-machine replication for planet-scale systems

why replicate?

- fault-tolerance

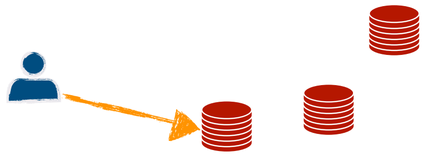


state-machine replication for planet-scale systems

why replicate?

- fault-tolerance

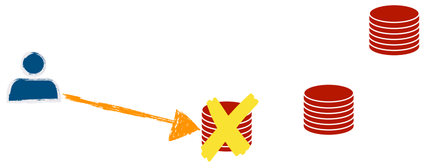
state-machine replication for planet-scale systems



why replicate?

- fault-tolerance

state-machine replication for planet-scale systems



why replicate?

- fault-tolerance

state-machine replication for planet-scale systems

why replicate?

- fault-tolerance

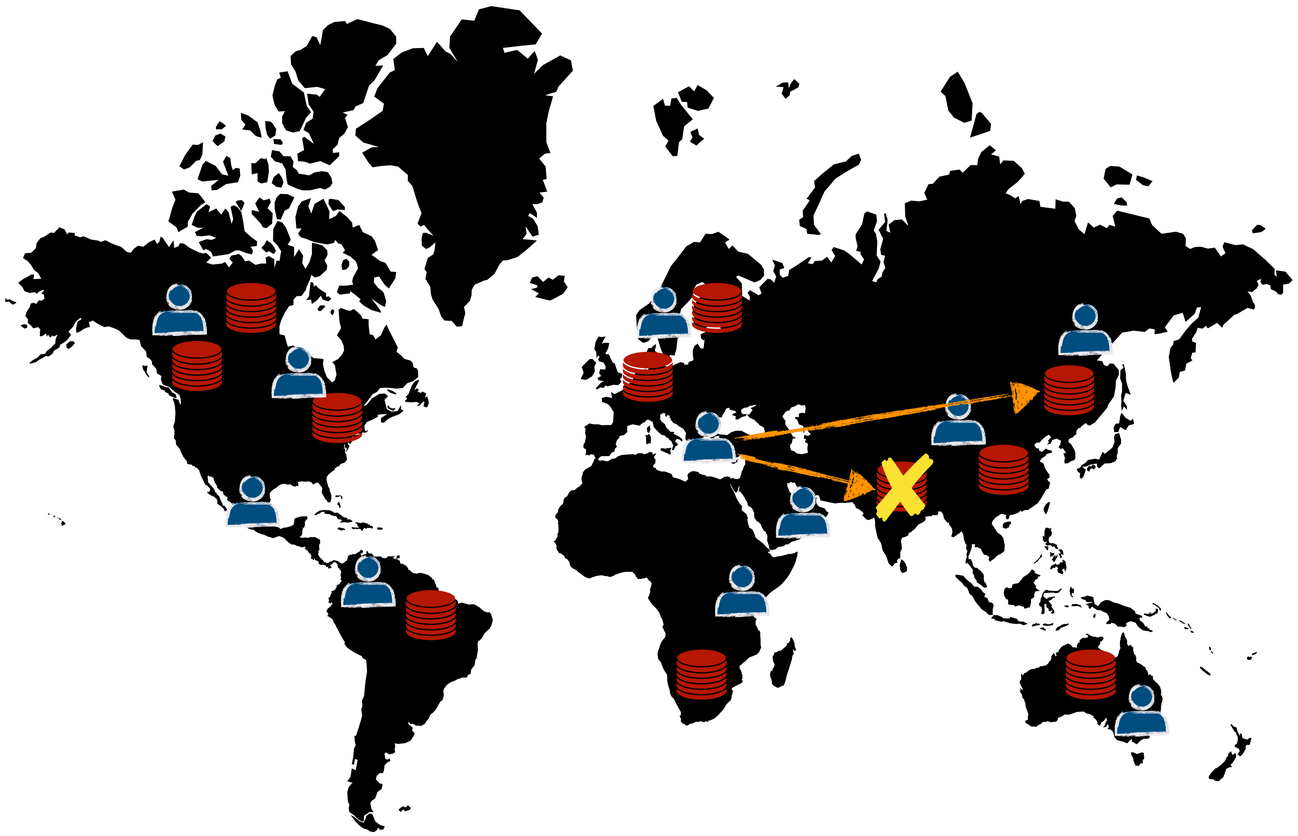
state-machine replication for planet-scale systems

why replicate?

- fault-tolerance

why planet-wide?

- minimizes latency



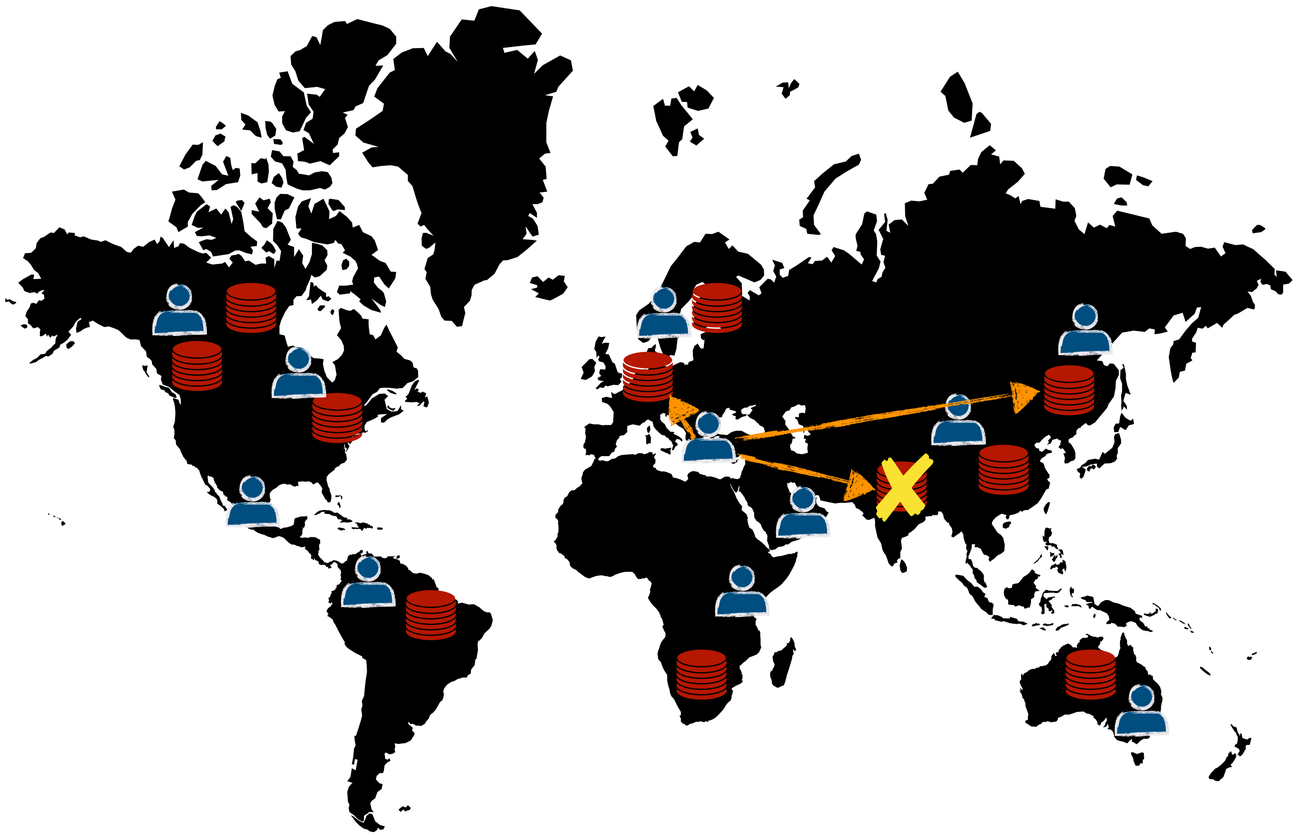
state-machine replication for planet-scale systems

why replicate?

- fault-tolerance

why planet-wide?

- minimizes latency



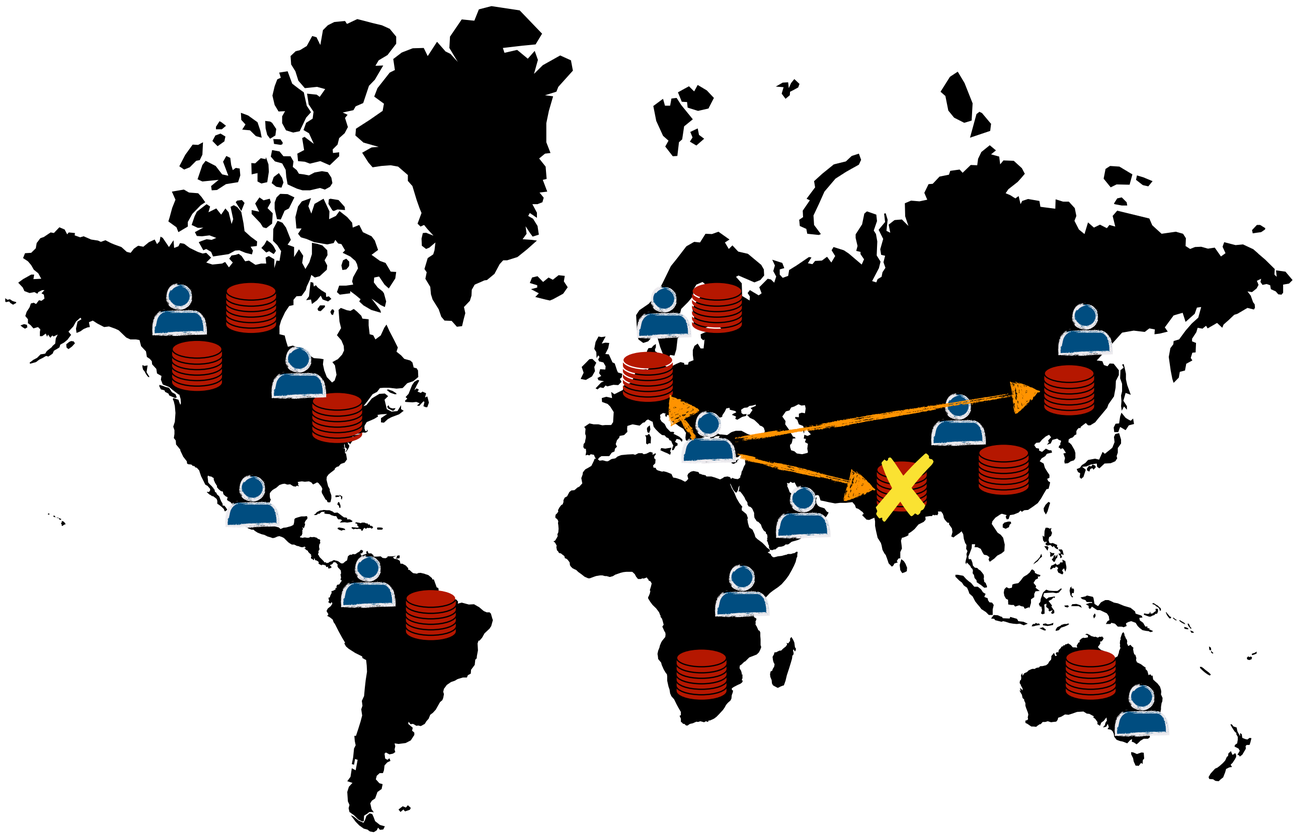
state-machine replication for planet-scale systems

why replicate?

- fault-tolerance

why planet-wide?

- minimizes latency



state-machine replication for planet-scale systems

why replicate?

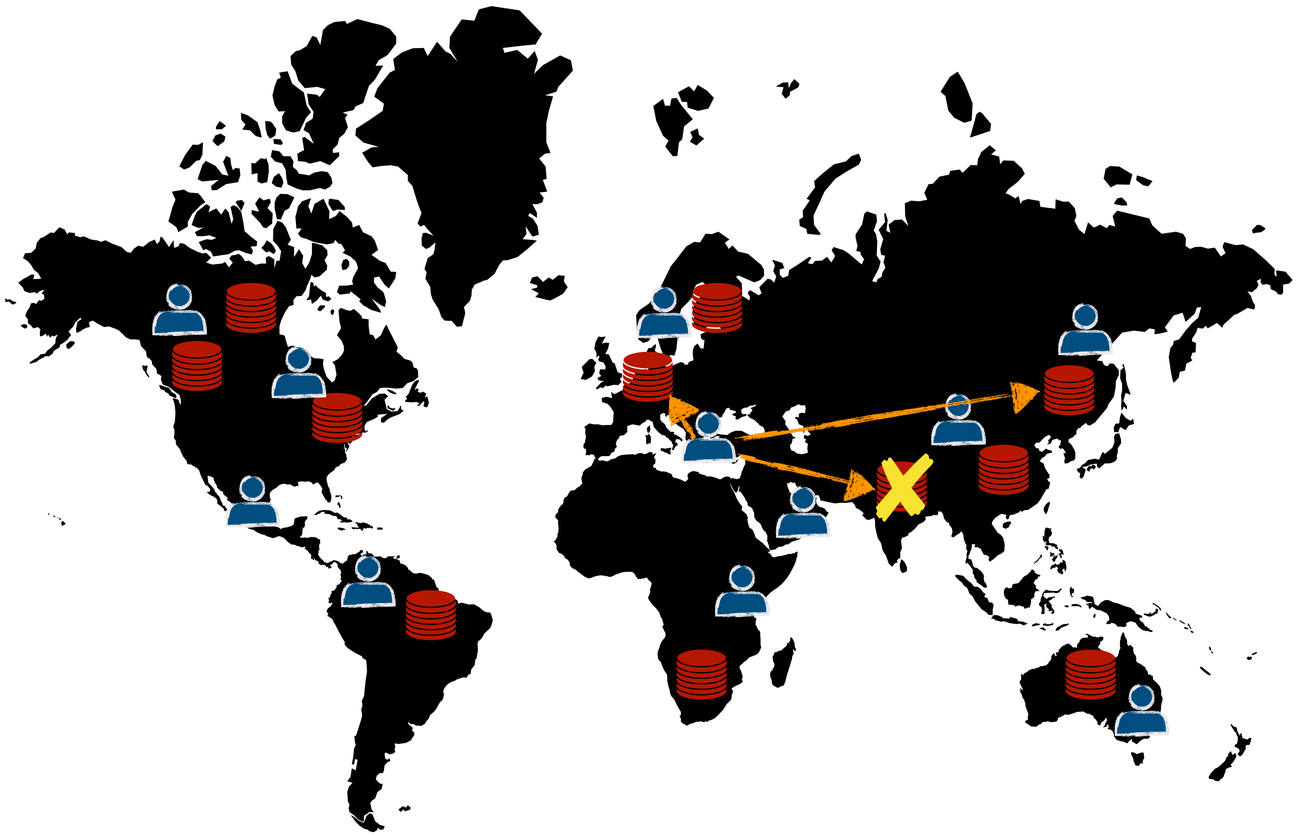
- fault-tolerance

why planet-wide?

- minimizes latency

how consistent?

- linearizable *(as single-copy)*



state-machine replication for planet-scale systems

why replicate?

- fault-tolerance

why planet-wide?

- minimizes latency

how consistent?

- linearizable *(as single-copy)*

how?

- state-machine replication (SMR)

can we attain low-latency with planet-scale SMR?

can we attain low-latency with planet-scale SMR?

small (nearby) quorums

leaderless

single round-trip

can we attain low-latency with planet-scale SMR?

small (nearby) quorums

leaderless

- no single point of failure
- load balancing
- no additional hops to the leader

single round-trip

SMR protocols

	small quorums	leaderless	single round-trip
(flexible) paxos			
epaxos			

SMR protocols

f: number of allowed failures

	small quorums	leaderless	single round-trip
(flexible) paxos	 f+1		 only from the leader
epaxos			

SMR protocols

f: number of allowed failures

	small quorums	leaderless	single round-trip
(flexible) paxos	✓ f+1	✗	− only from the leader
epaxos	✗ 3n/4	✓	− only if no conflicts

SMR protocols

f: number of allowed failures

	small quorums	leaderless	single round-trip
(flexible) paxos	✓ f+1	✗	− only from the leader
epaxos	✗ 3n/4	✓	− only if no conflicts
atlas (this paper)			

SMR protocols

f: number of allowed failures

	small quorums	leaderless	single round-trip
(flexible) paxos	✓ f+1	✗	− only from the leader
epaxos	✗ 3n/4	✓	− only if no conflicts
atlas (this paper)			

3-month link-monitoring experiment

observation: concurrent link slowdowns are rare!

(high values for f are unnecessary)

SMR protocols

f: number of allowed failures

	small quorums	leaderless	single round-trip
(flexible) paxos	✓ f+1	✗	− only from the leader
epaxos	✗ 3n/4	✓	− only if no conflicts
atlas (this paper)	✓ n/2+f	✓	✓

3-month link-monitoring experiment

observation: concurrent link slowdowns are rare!

(high values for f are unnecessary)

SMR protocols

f: number of allowed failures

	small quorums	leaderless	single round-trip
(flexible) paxos	✓ f+1	✗	− only from the leader
epaxos	✗ 3n/4	✓	− only if no conflicts
atlas (this paper)	✓ n/2+f	✓	✓

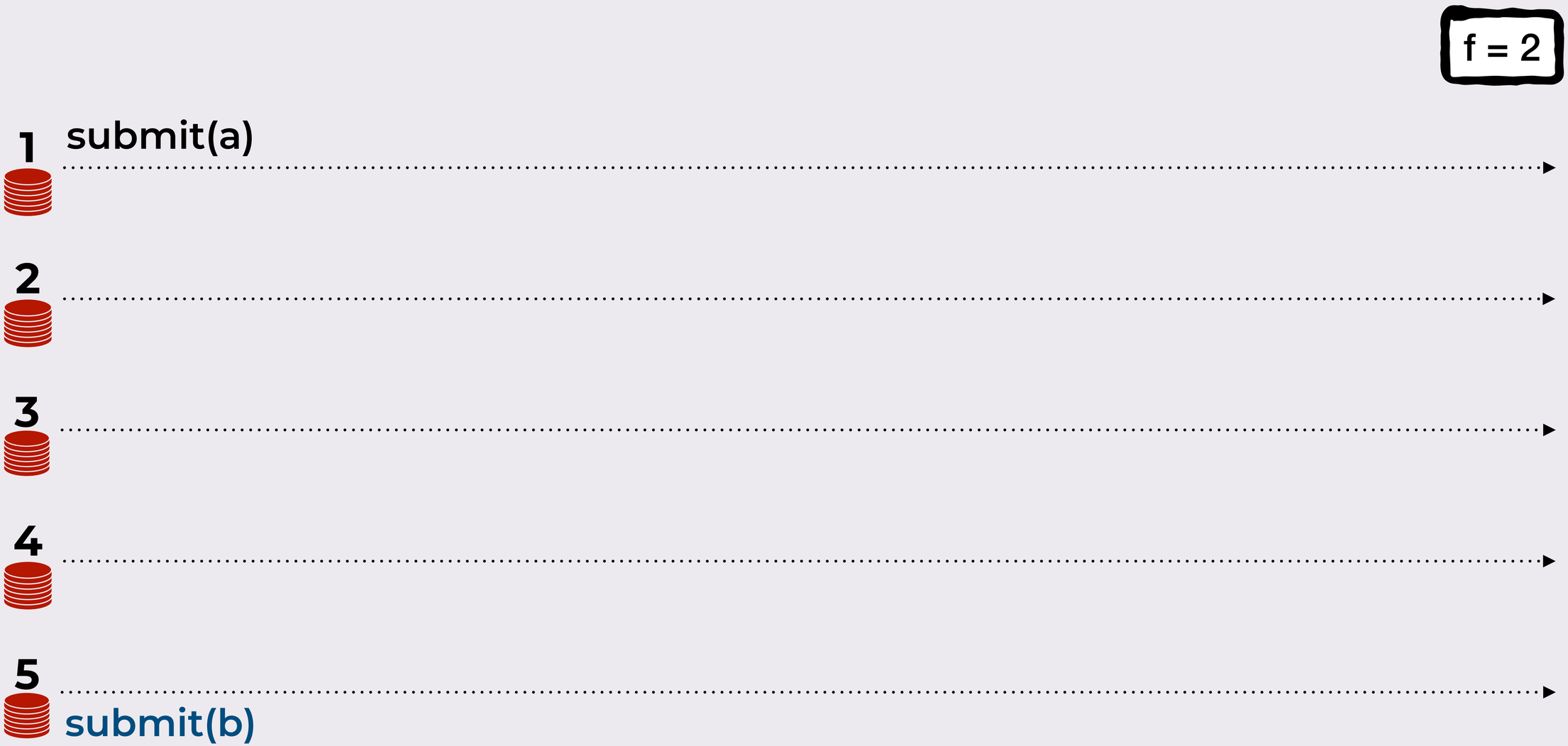
flexible fast-path condition!

3-month link-monitoring experiment

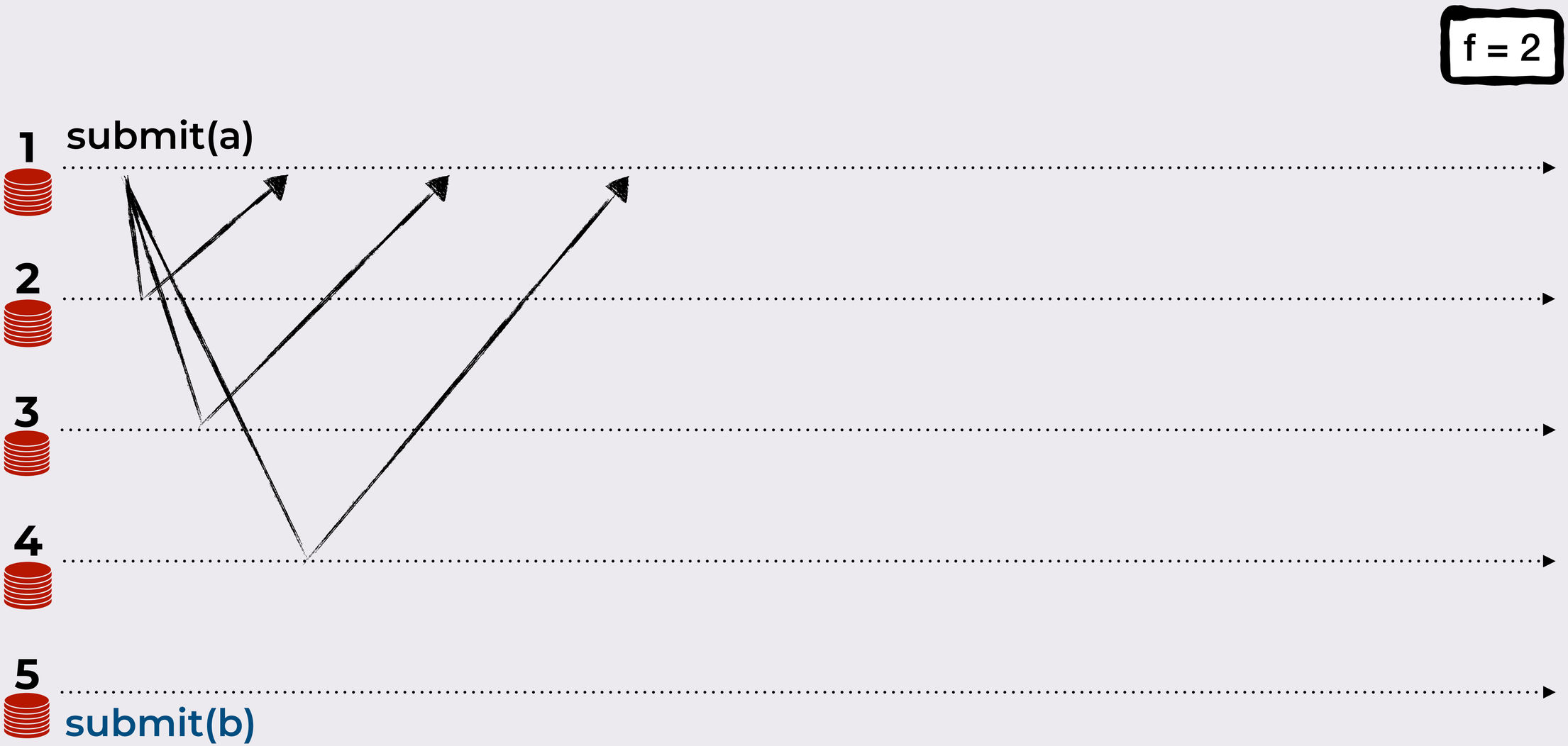
observation: concurrent link slowdowns are rare!

(high values for f are unnecessary)

flexible fast-path condition

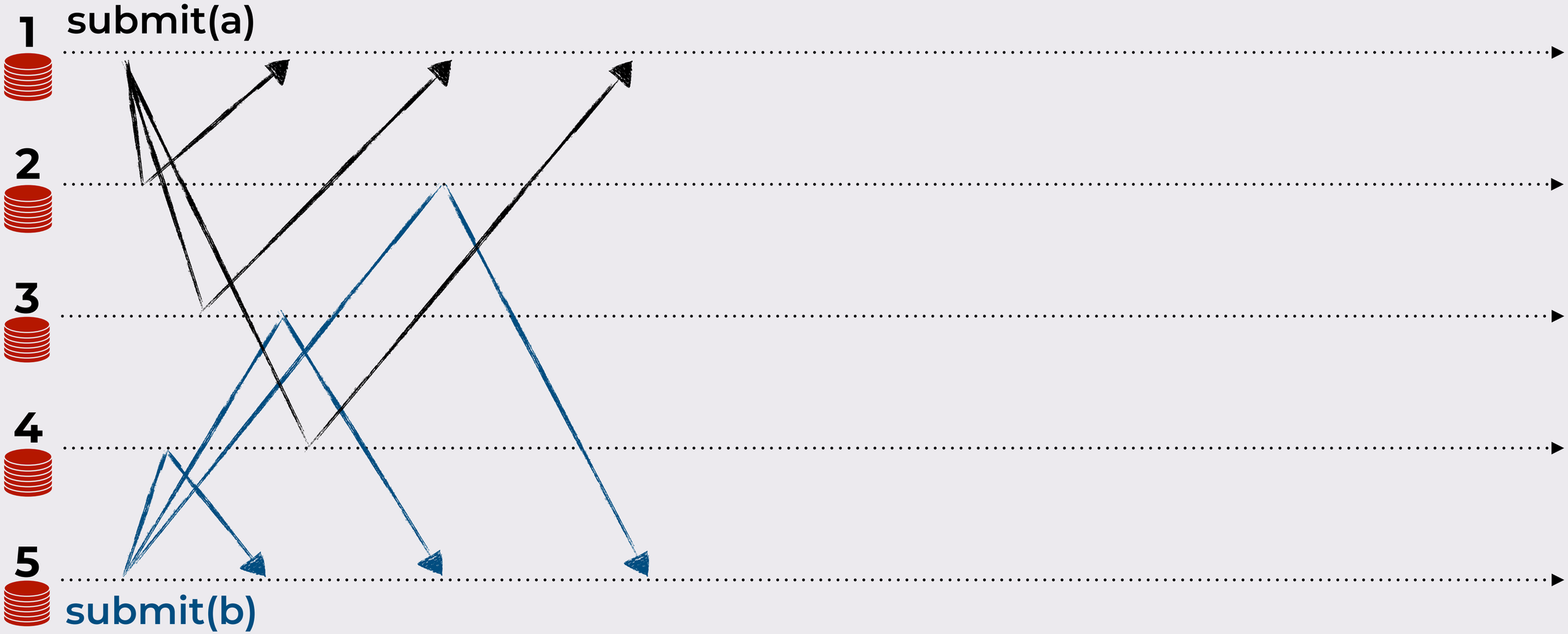


flexible fast-path condition

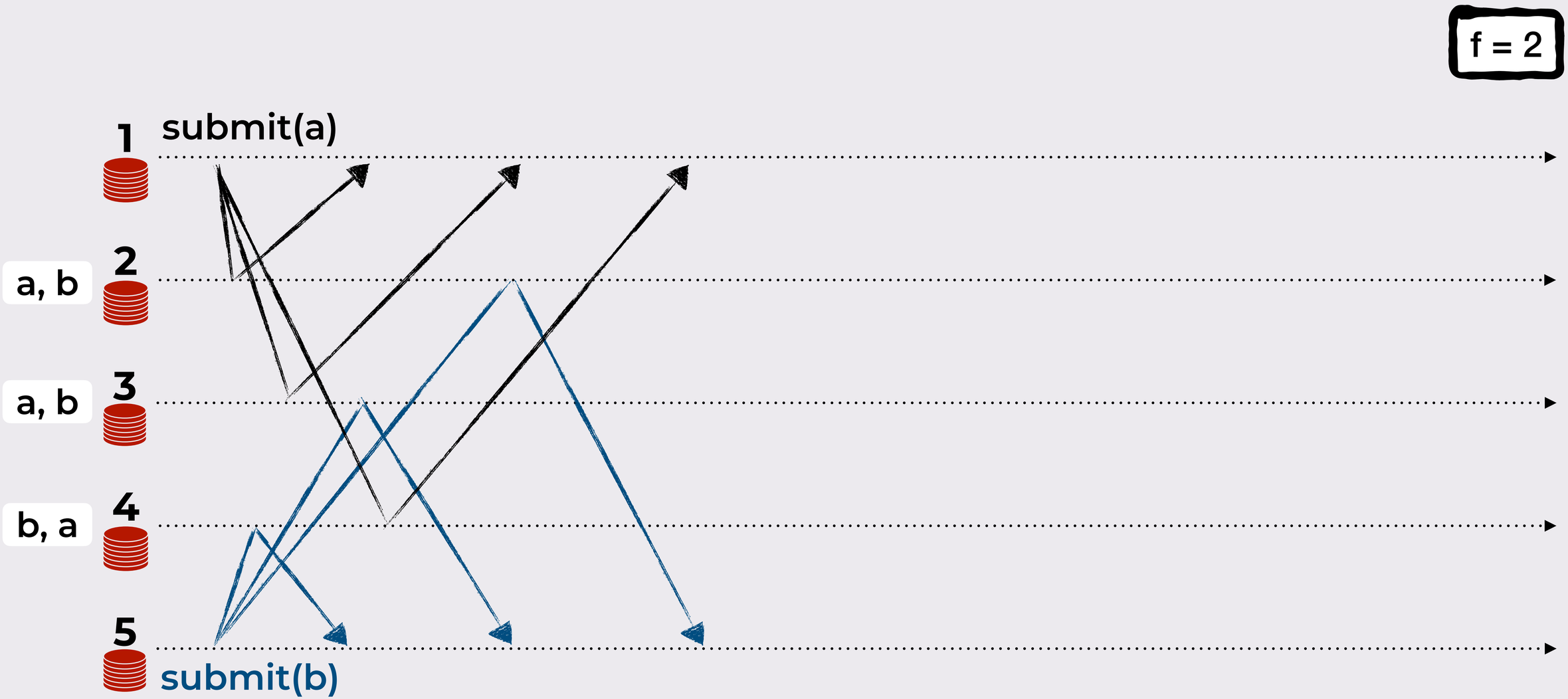


flexible fast-path condition

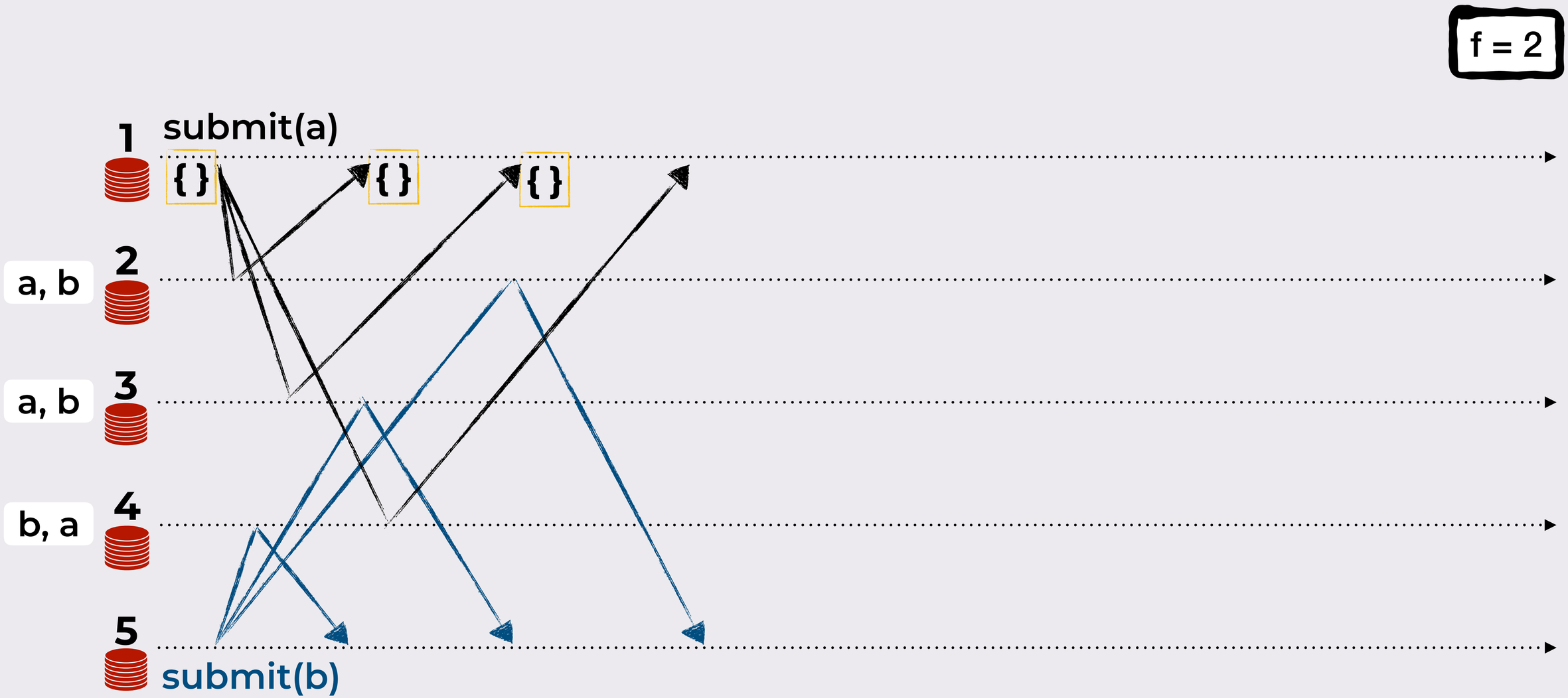
f = 2



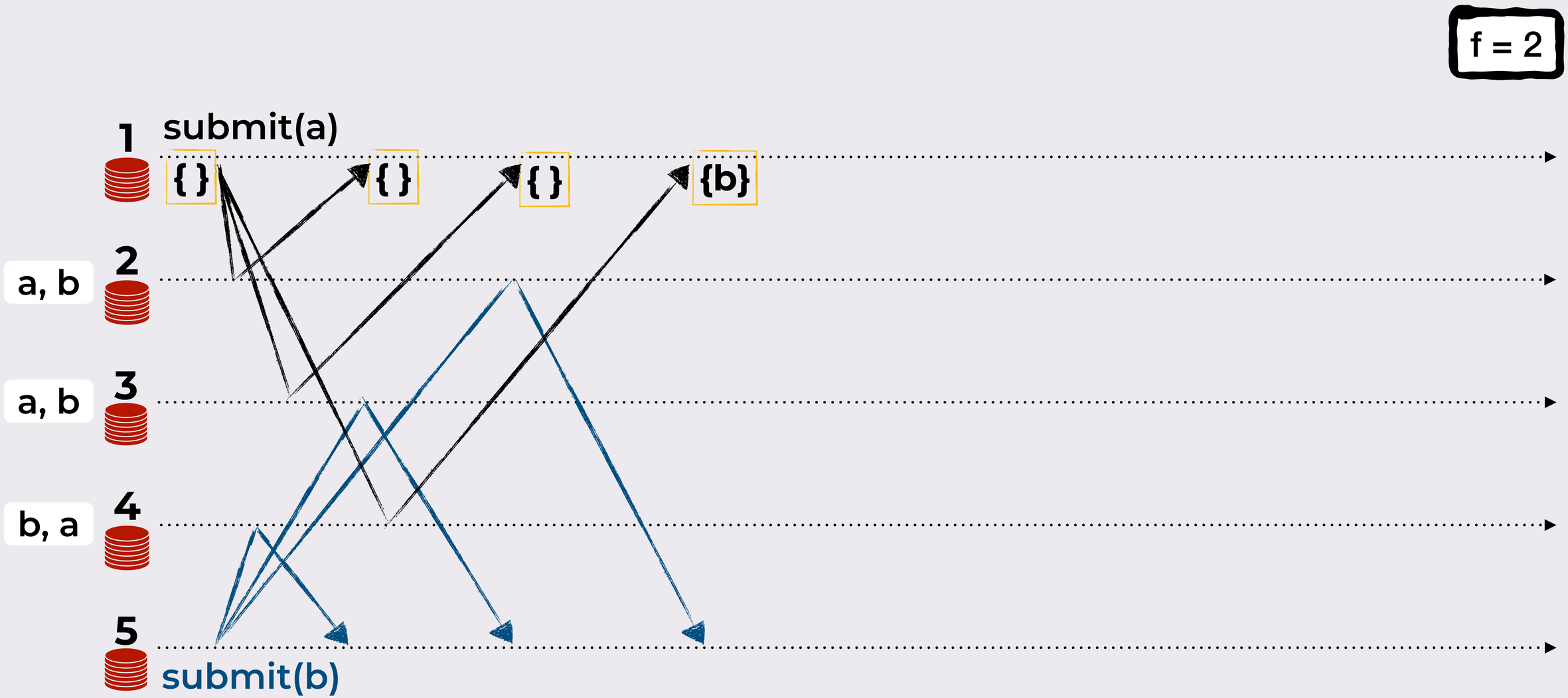
flexible fast-path condition



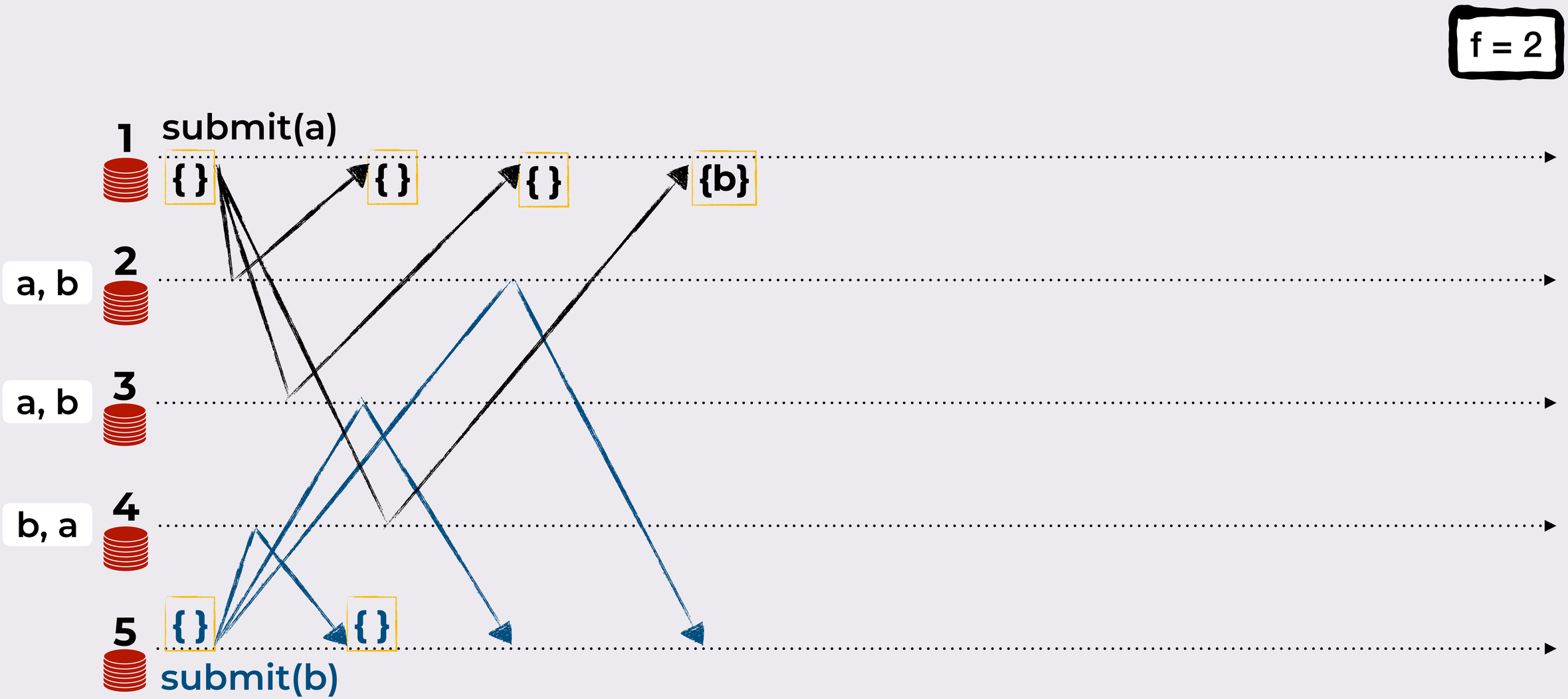
flexible fast-path condition



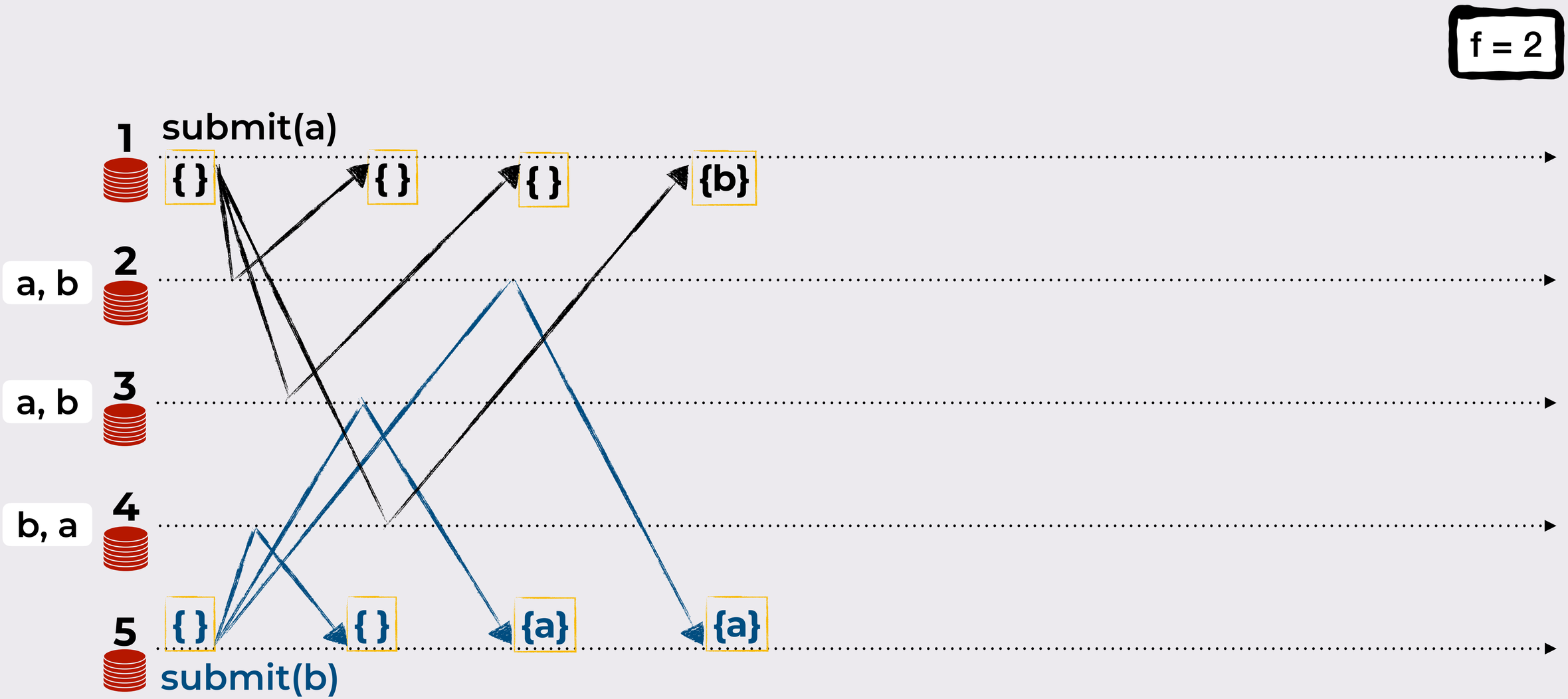
flexible fast-path condition



flexible fast-path condition

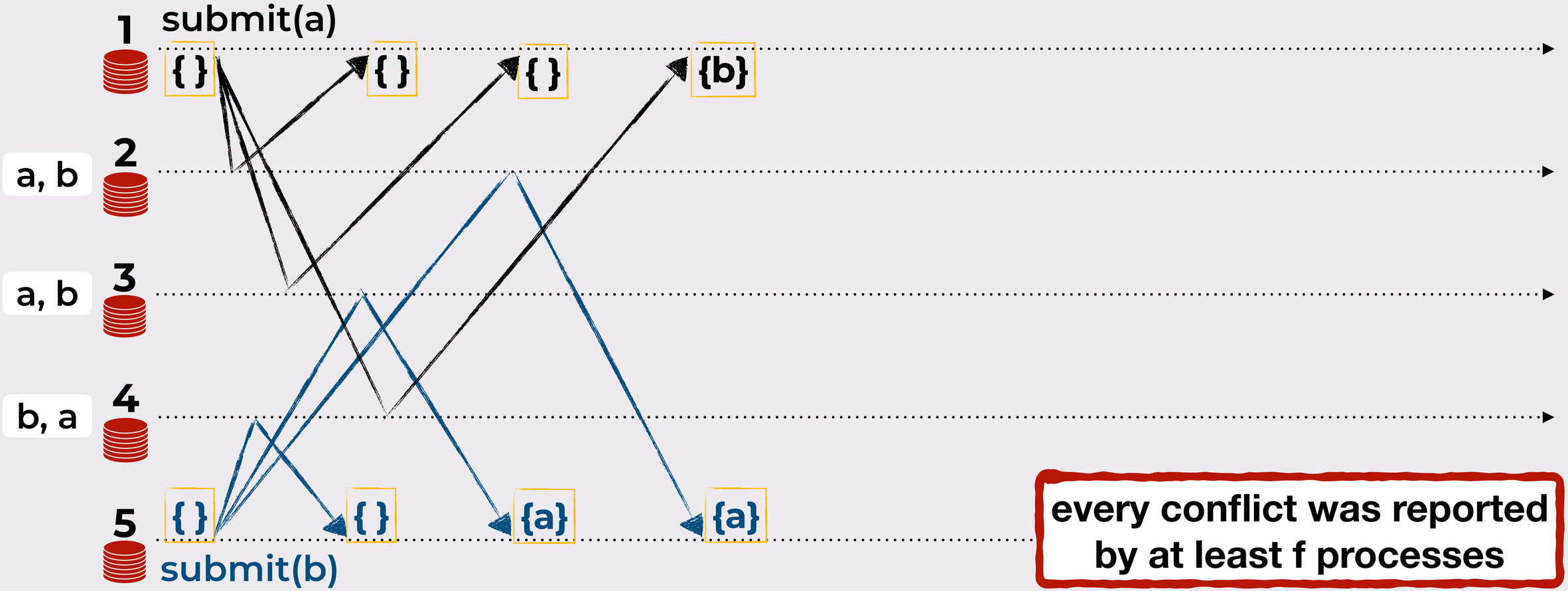


flexible fast-path condition



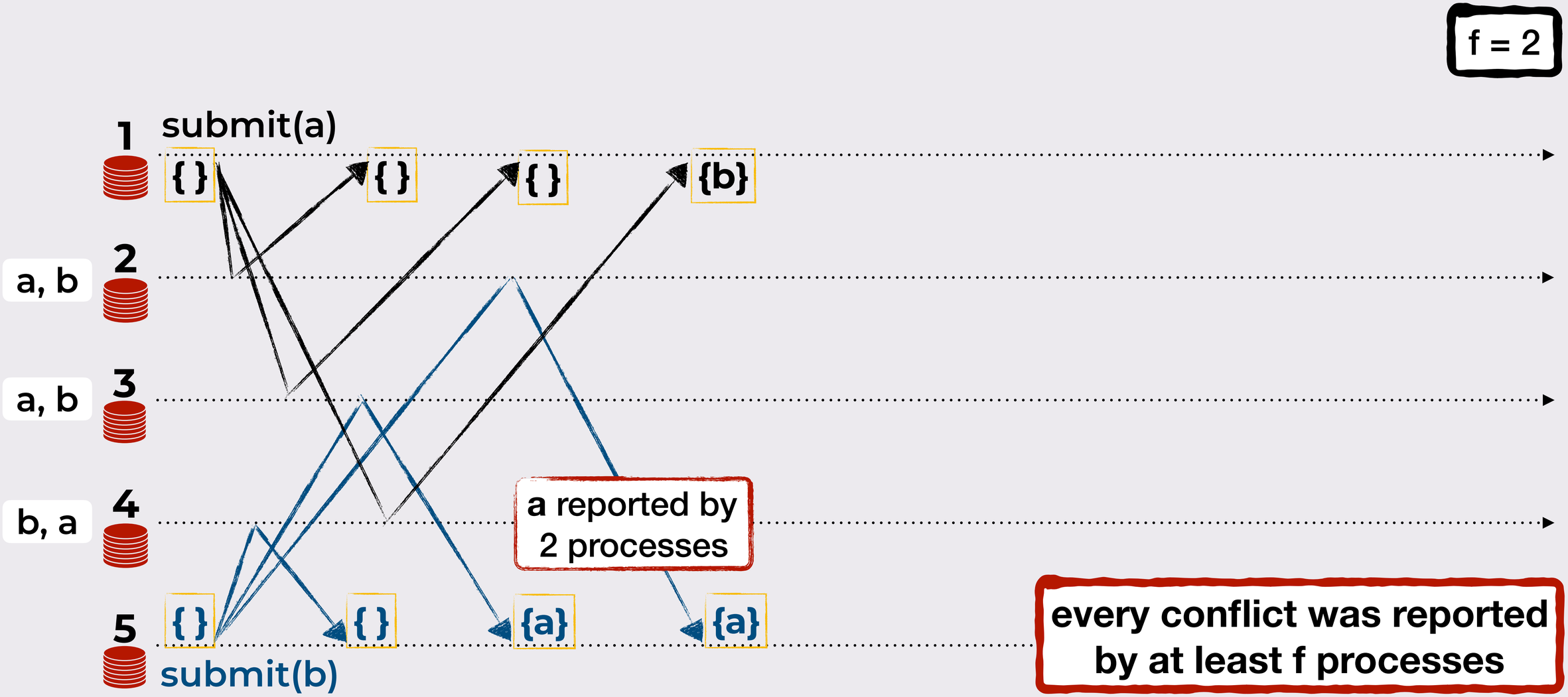
flexible fast-path condition

f = 2

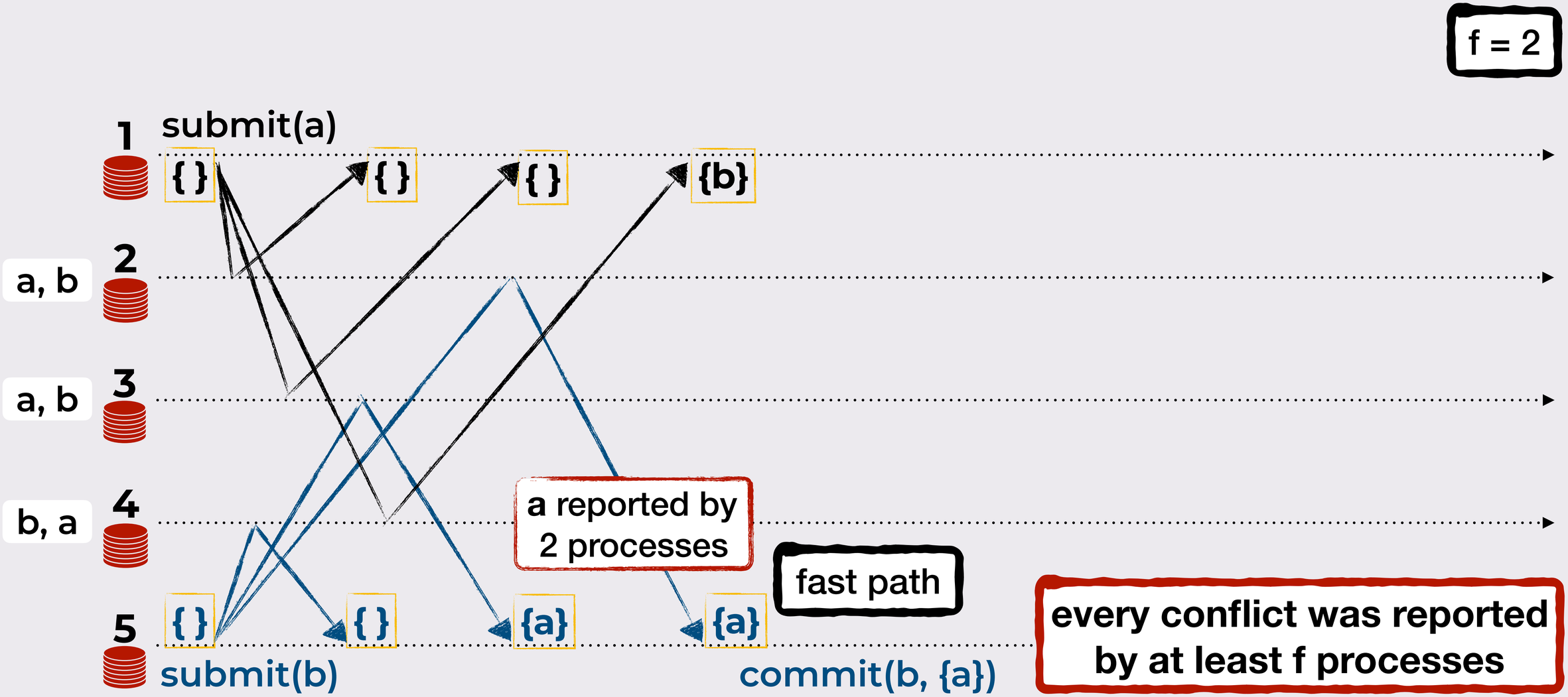


every conflict was reported by at least f processes

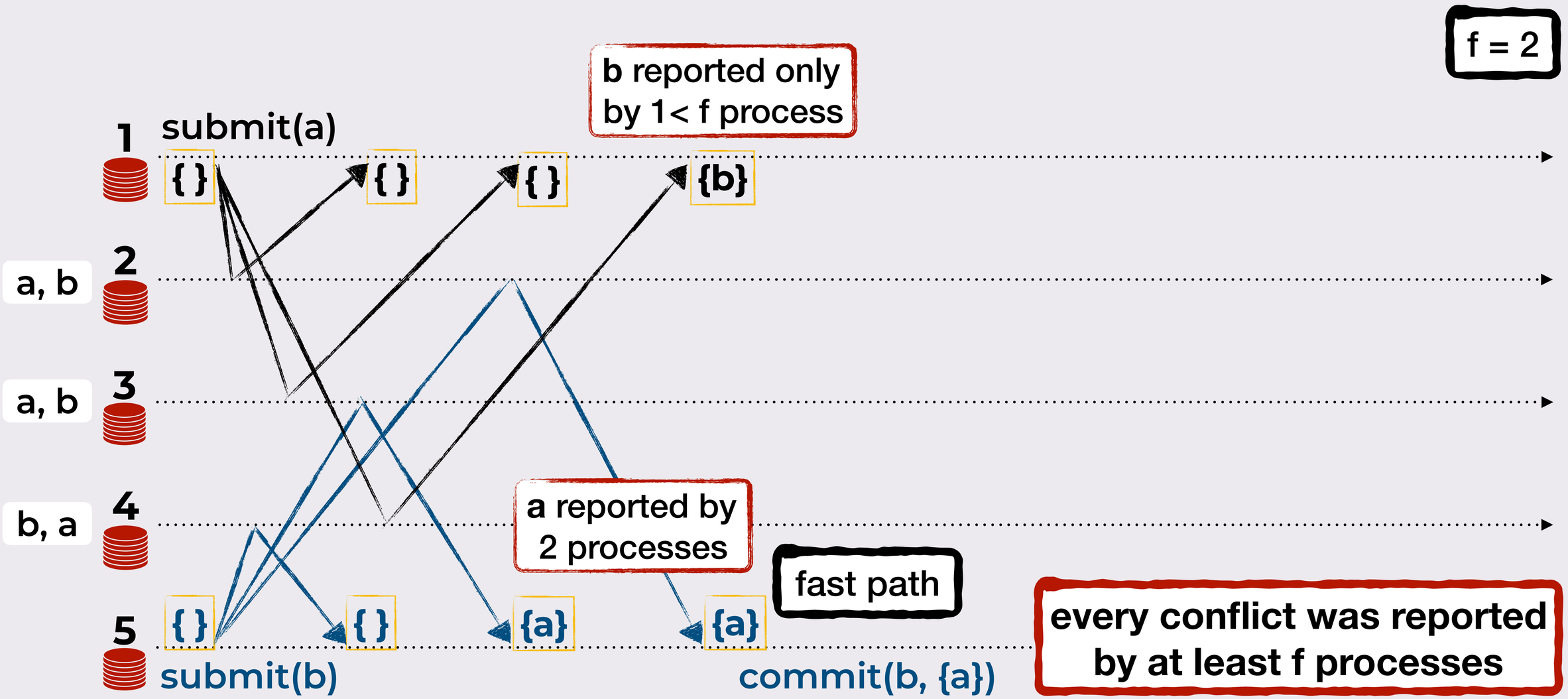
flexible fast-path condition



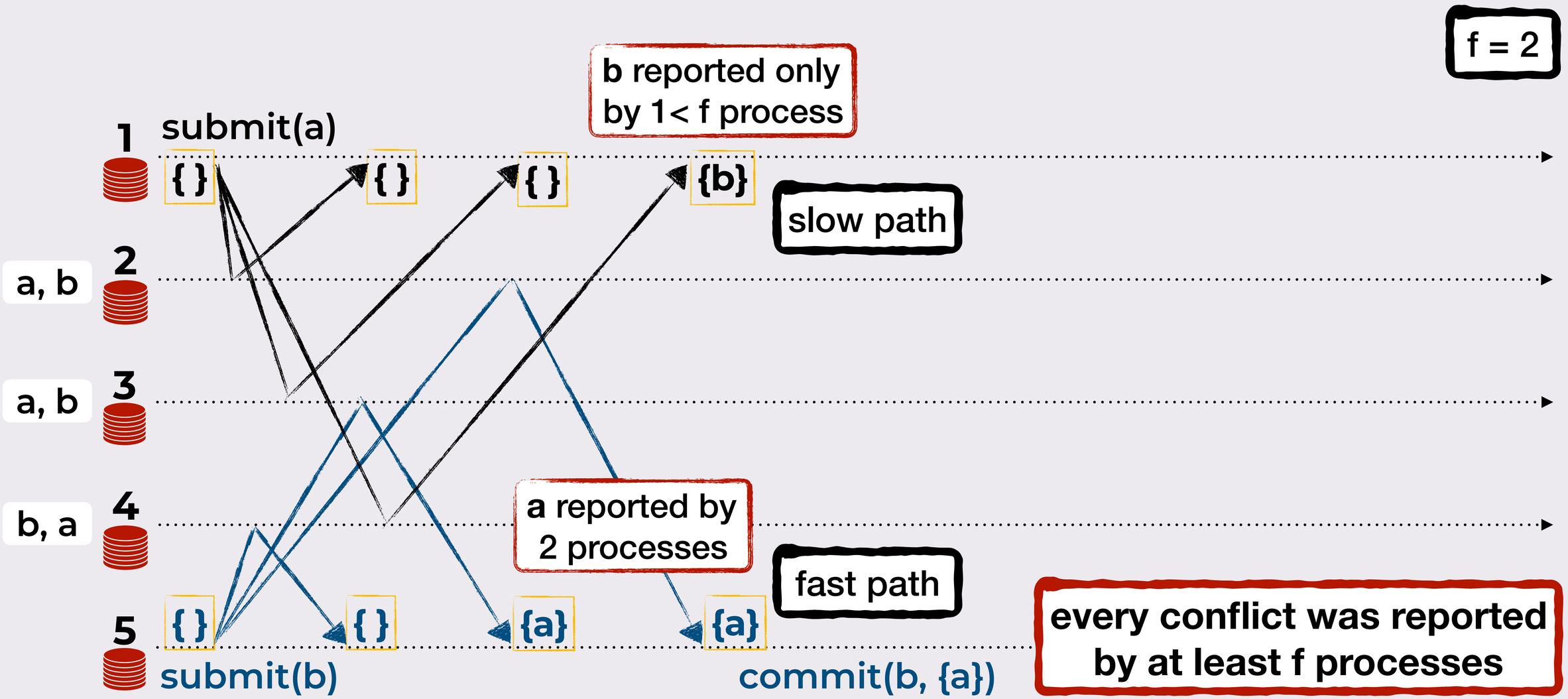
flexible fast-path condition



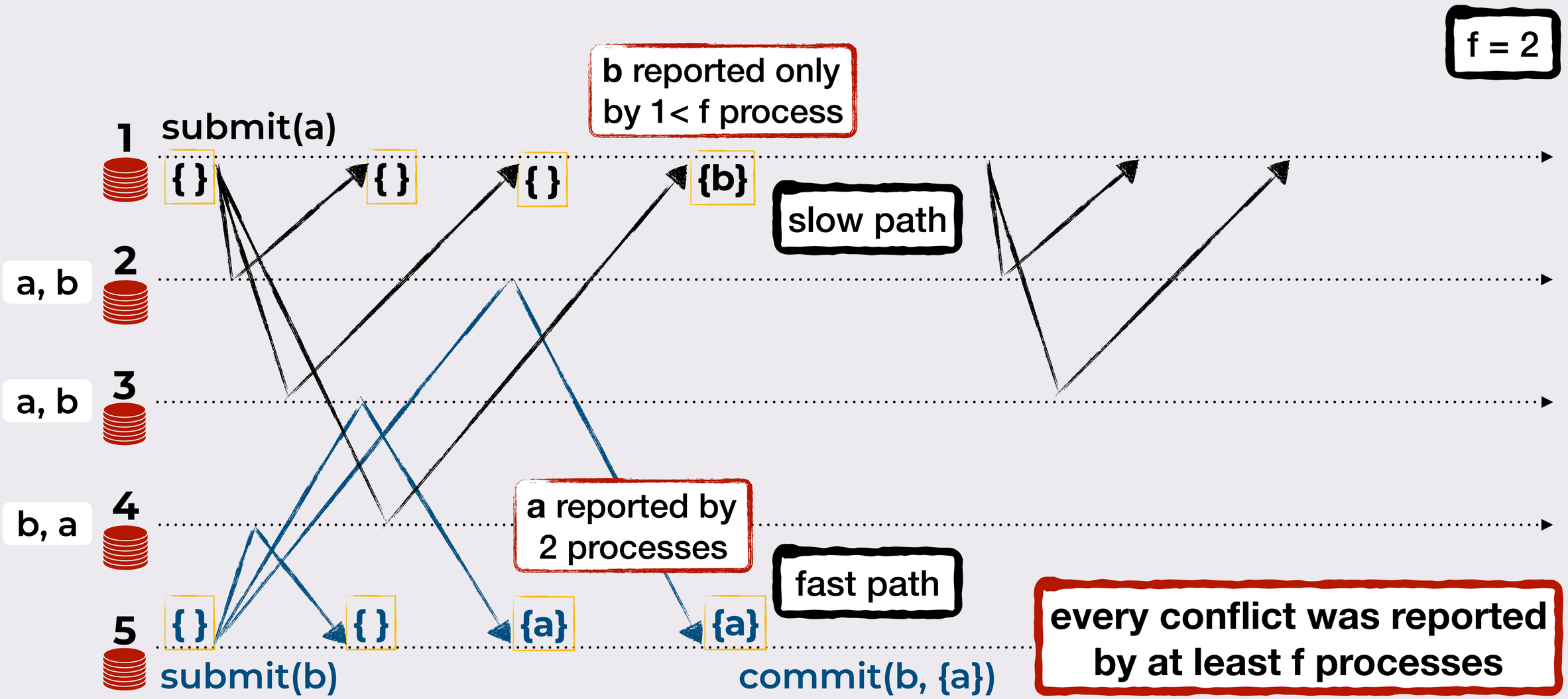
flexible fast-path condition



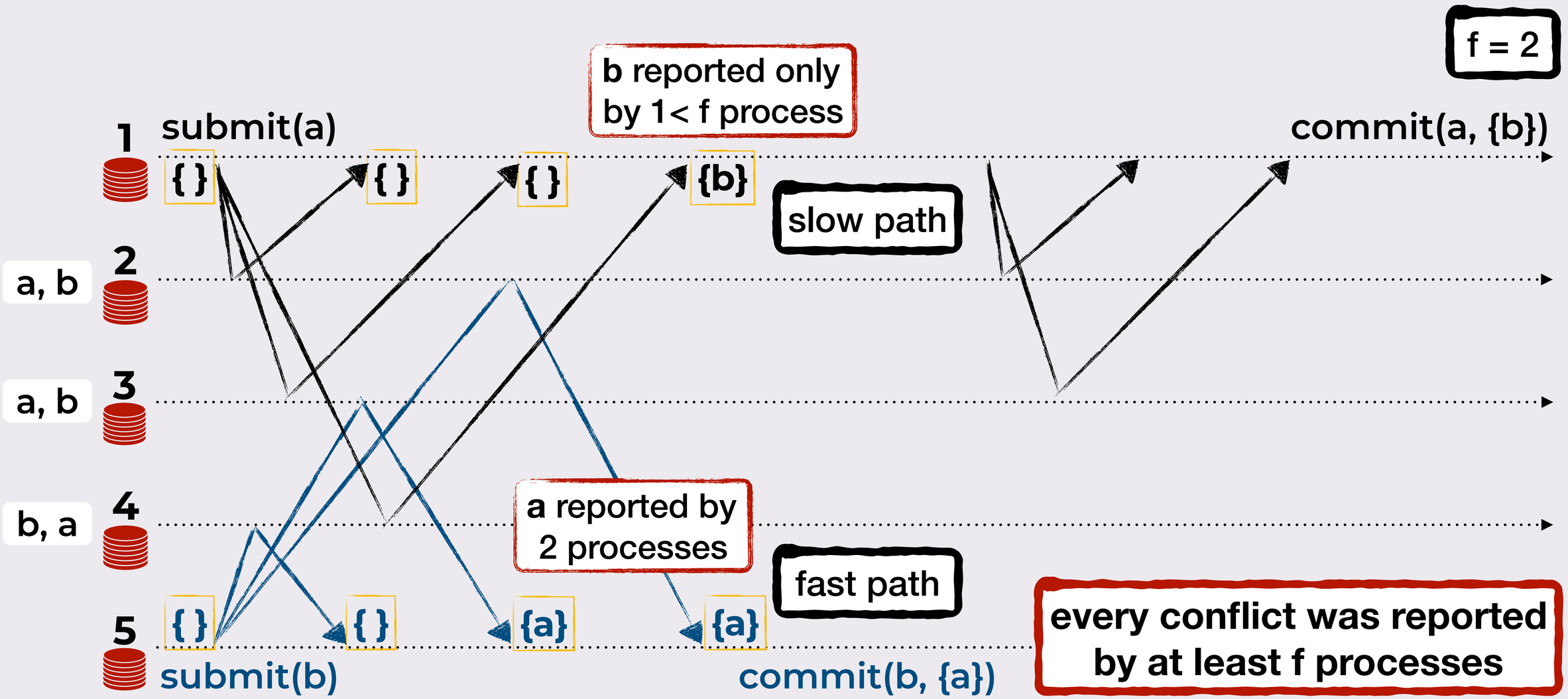
flexible fast-path condition



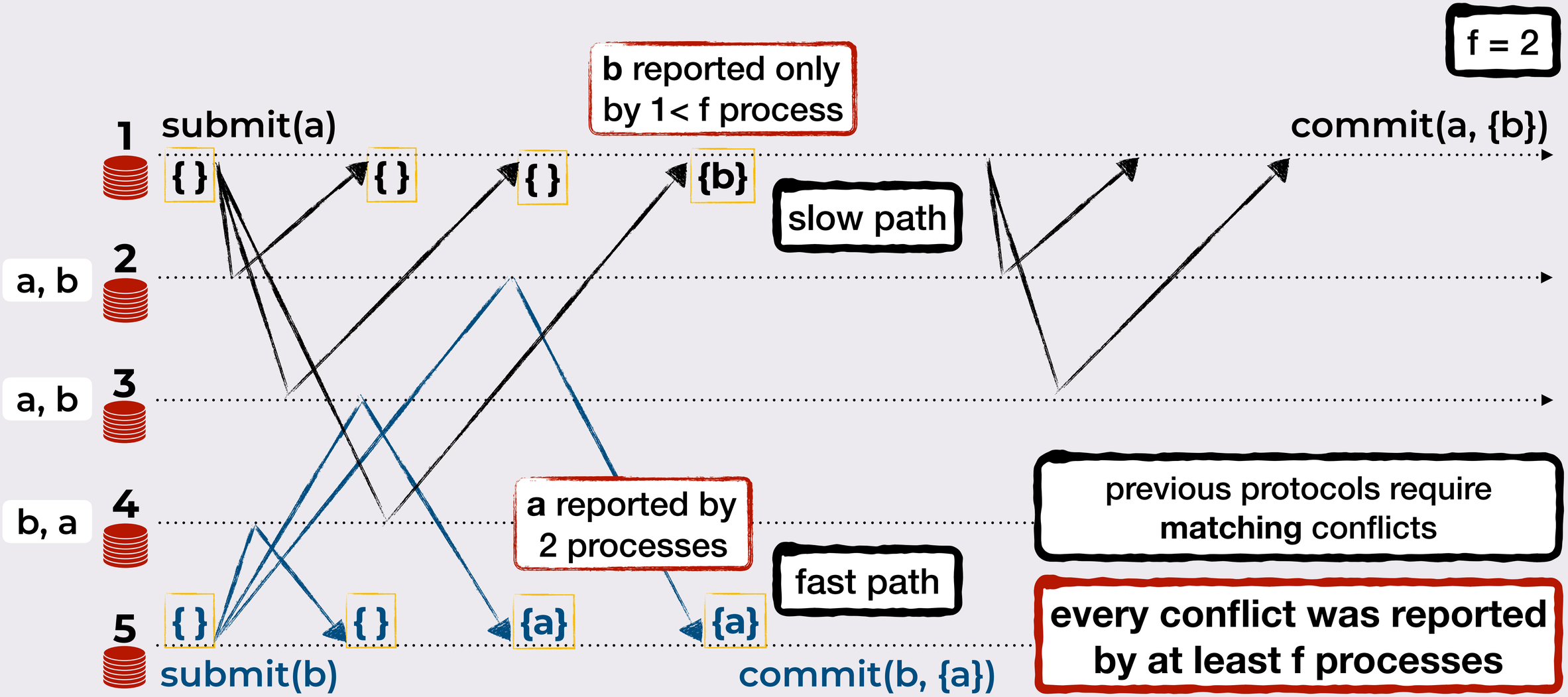
flexible fast-path condition



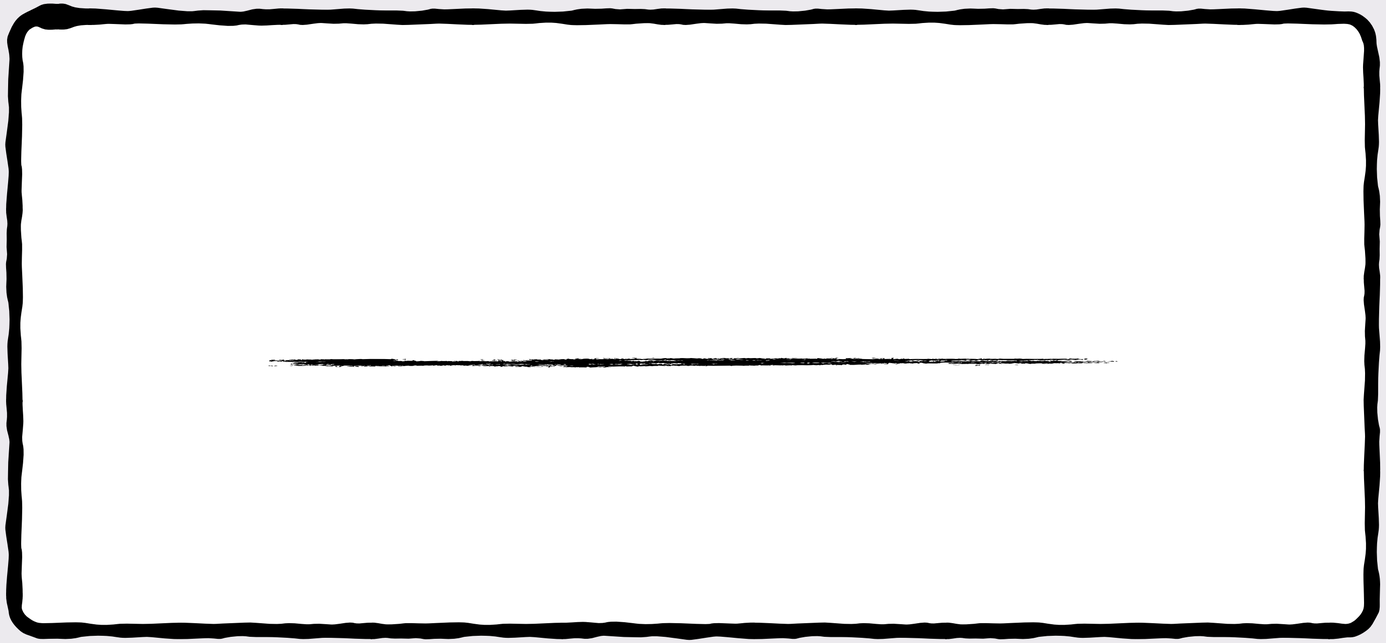
flexible fast-path condition



flexible fast-path condition



command execution



committed conflicts determine command execution order

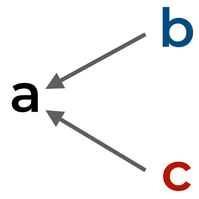
command execution

commit(a, { })
commit(b, {a, c})
commit(c, {a, b})

committed conflicts determine command execution order

command execution

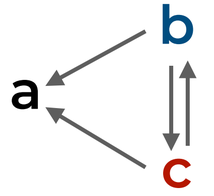
commit(a, { })
commit(b, {a, c})
commit(c, {a, b})



committed conflicts determine command execution order

command execution

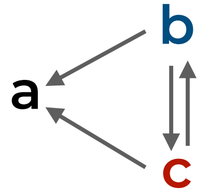
commit(a, { })
commit(b, {a, c})
commit(c, {a, b})



committed conflicts determine command execution order

command execution

commit(a, { })
commit(b, {a, c})
commit(c, {a, b})



---

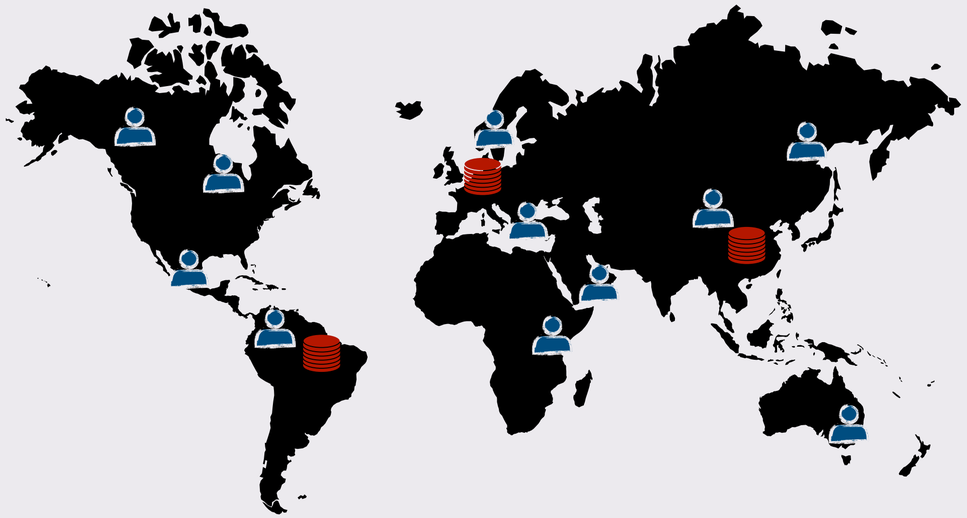
execute(a) ; execute(b) ; execute(c) if b < c
execute(a) ; execute(c) ; execute(b) if b > c

committed conflicts determine command execution order

planet-scale evaluation

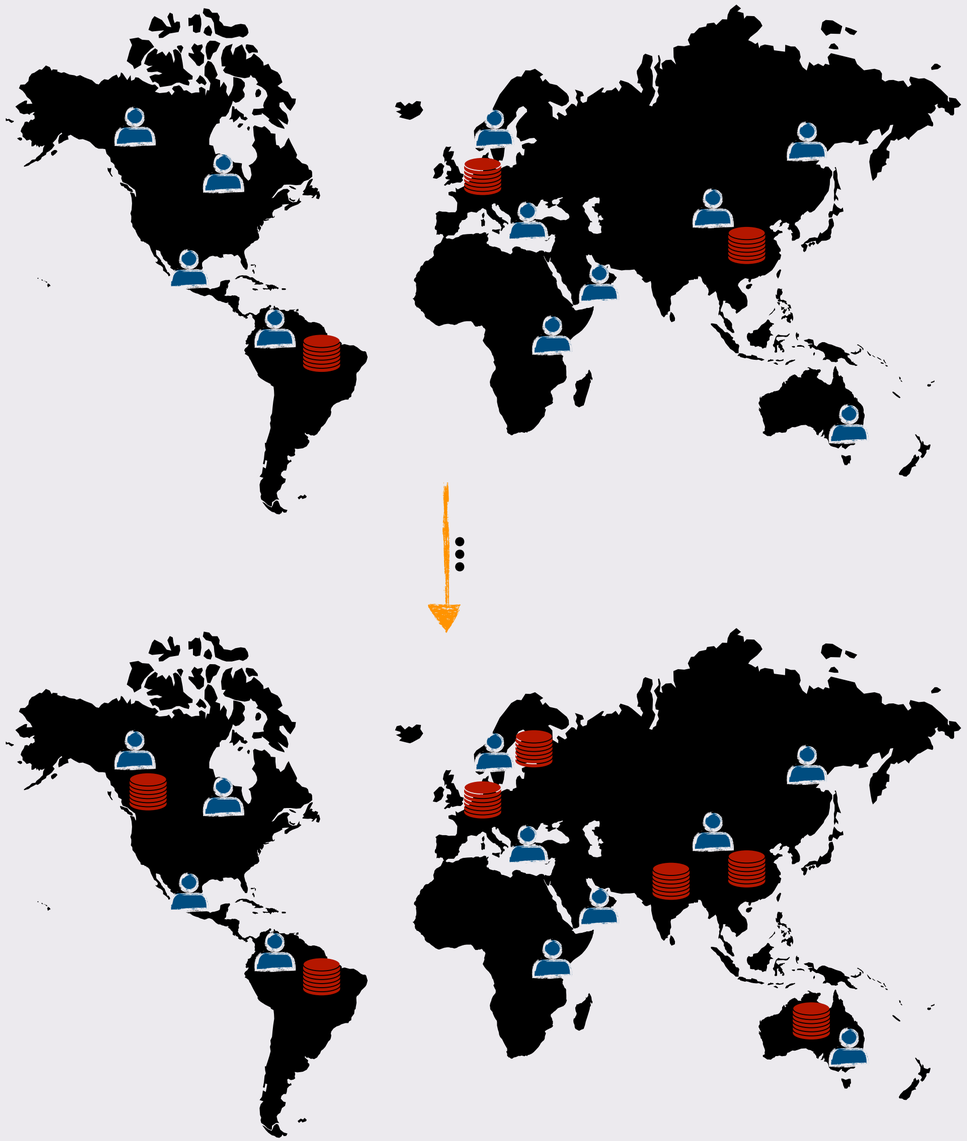
1. bringing the system closer to clients

planet-scale evaluation



1. bringing the system closer to clients

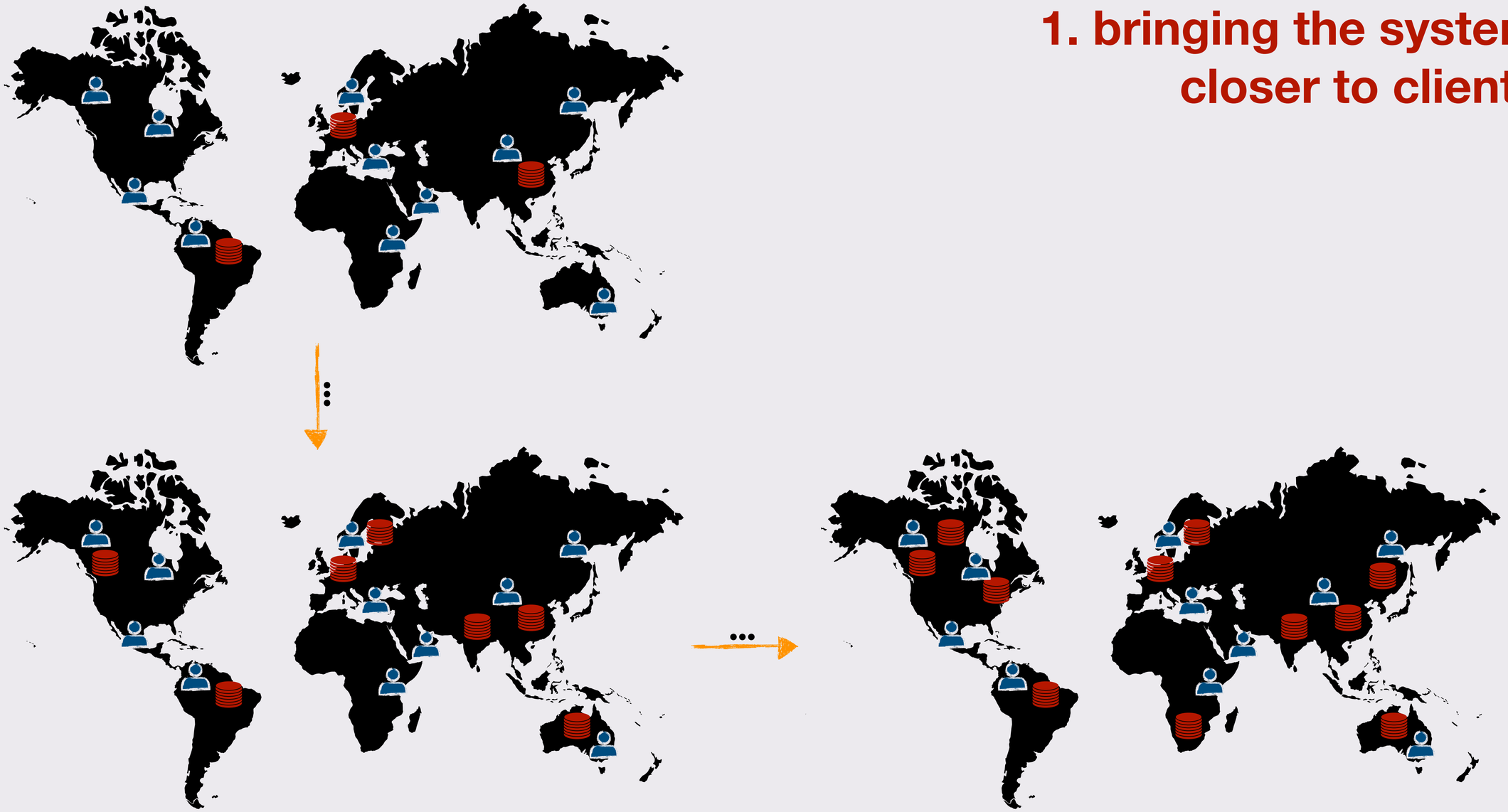
planet-scale evaluation



1. bringing the system closer to clients

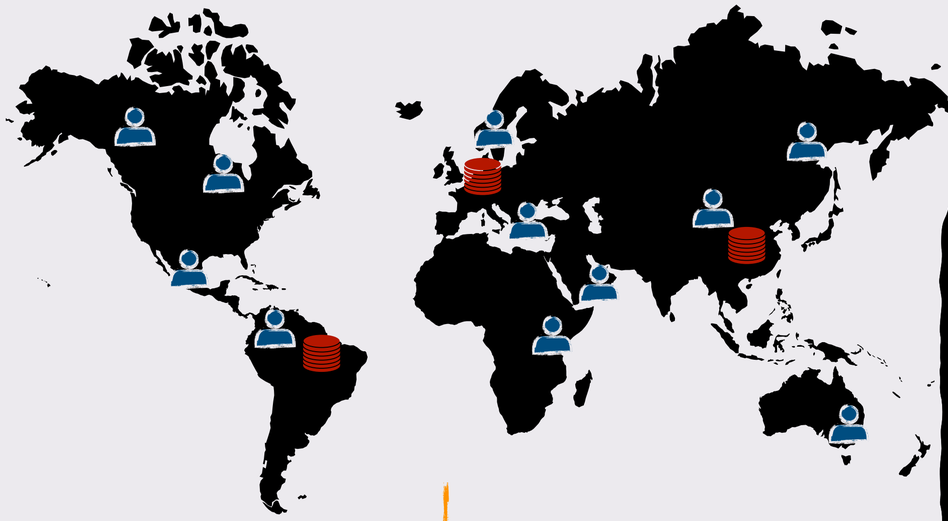
planet-scale evaluation

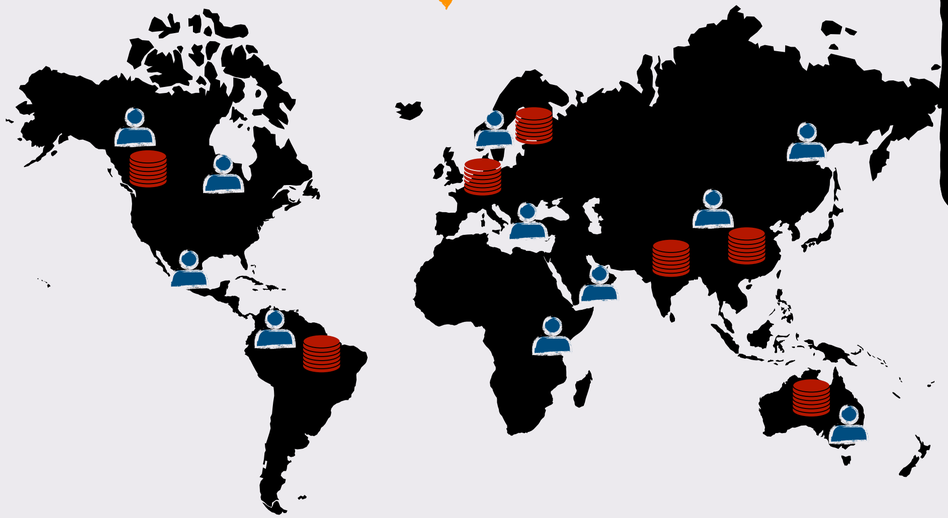
1. bringing the system closer to clients

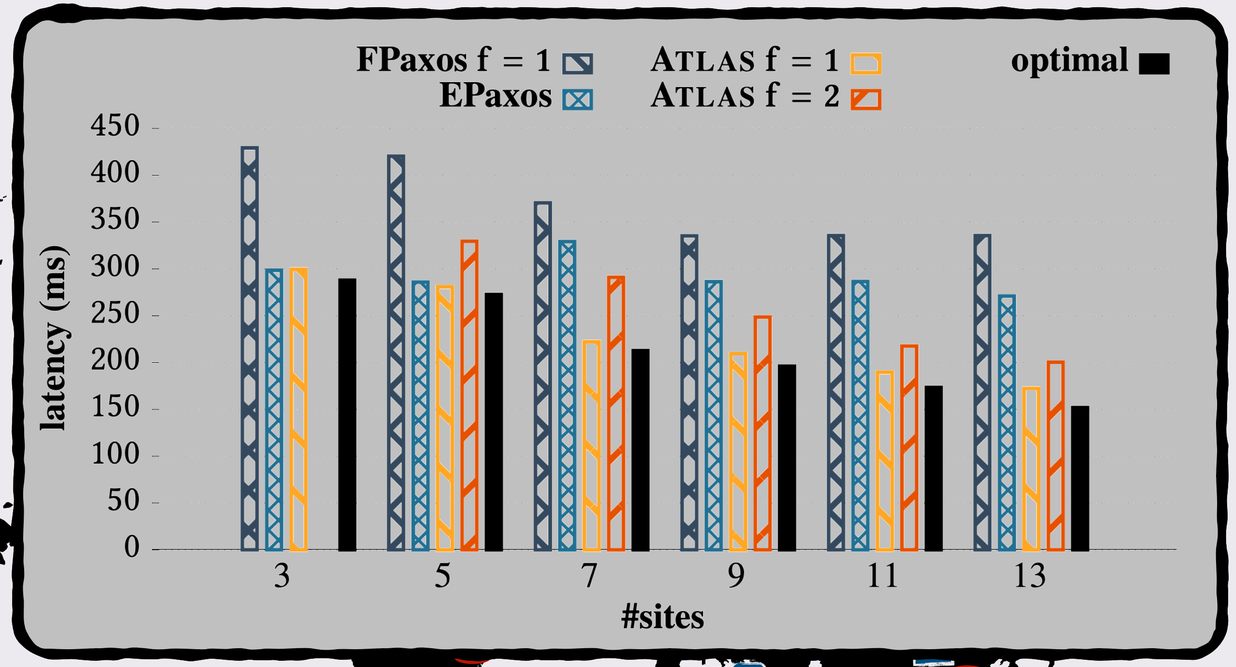


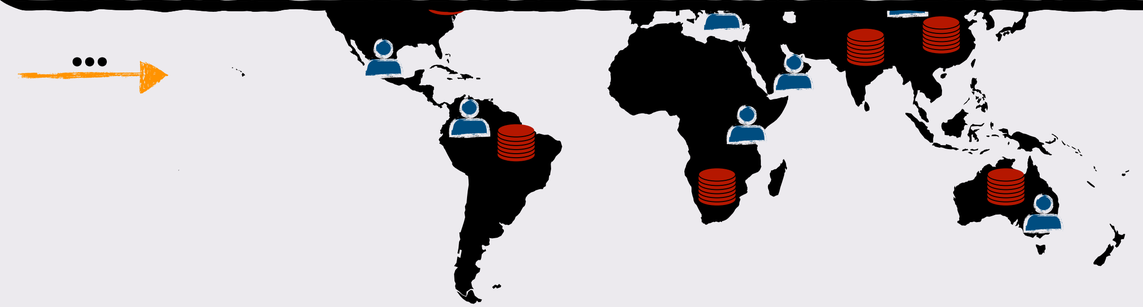
planet-scale evaluation

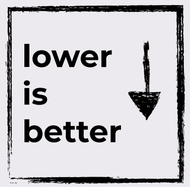
1. bringing the system closer to clients





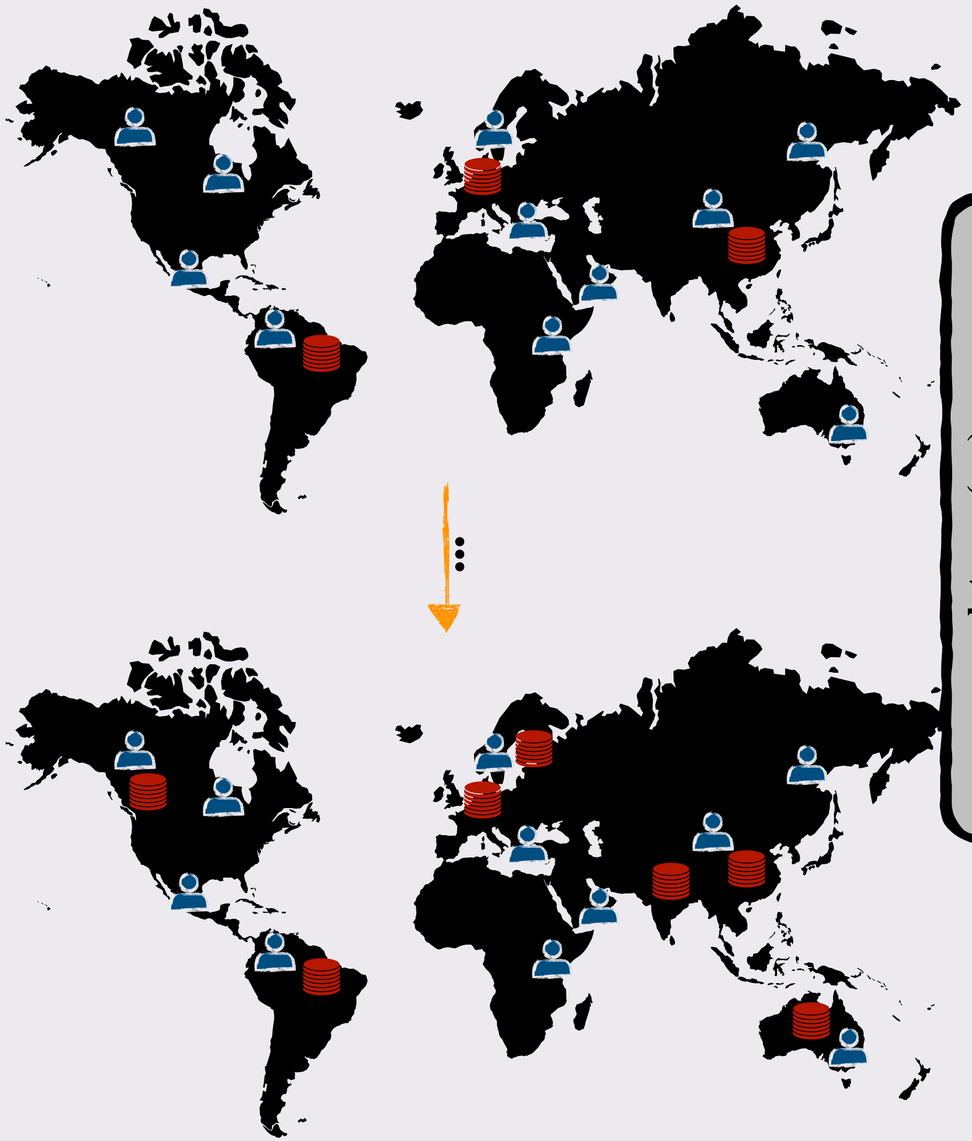


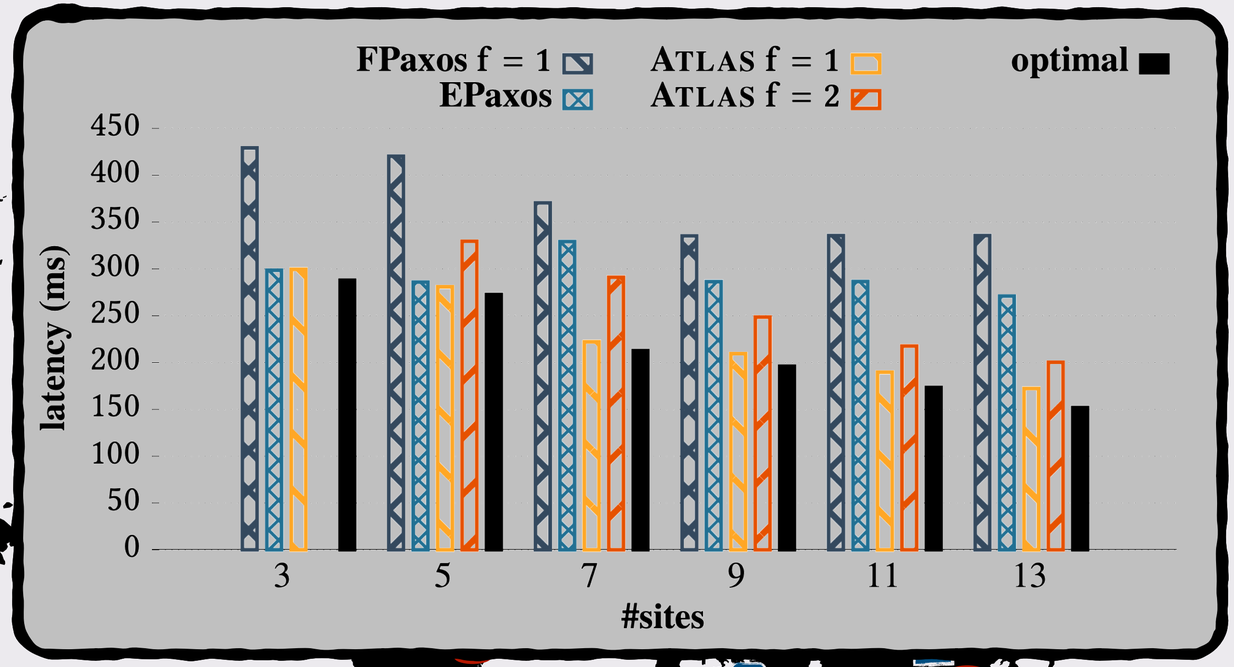




planet-scale evaluation

1. bringing the system closer to clients





lower is better

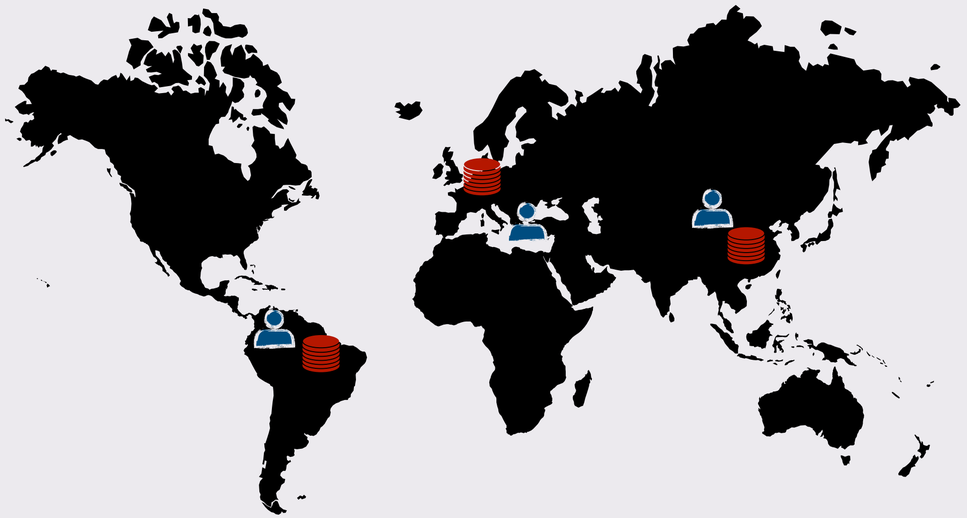
atlas improves performance as it is deployed closer to clients

planet-scale evaluation

2. expanding the system

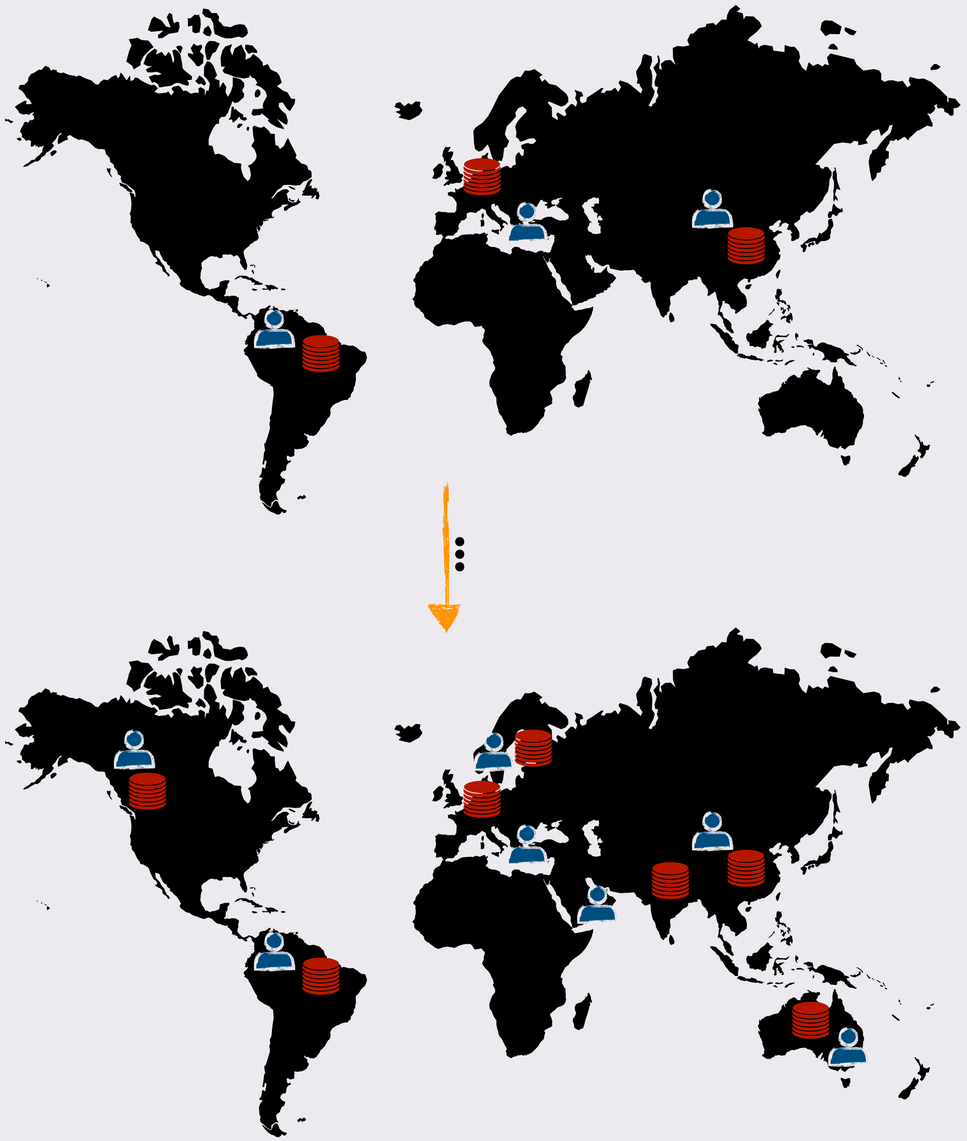
planet-scale evaluation

2. expanding the system



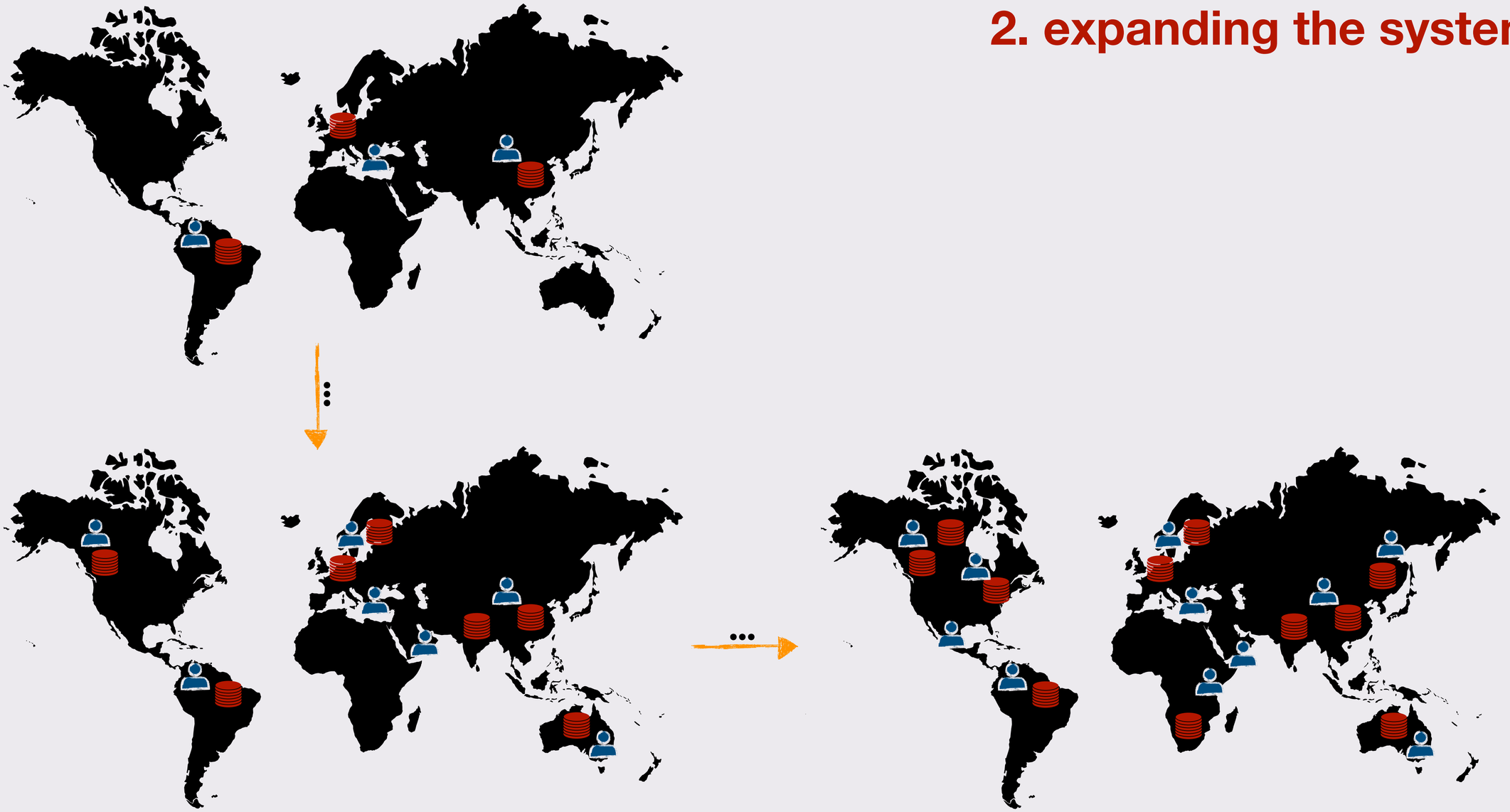
planet-scale evaluation

2. expanding the system



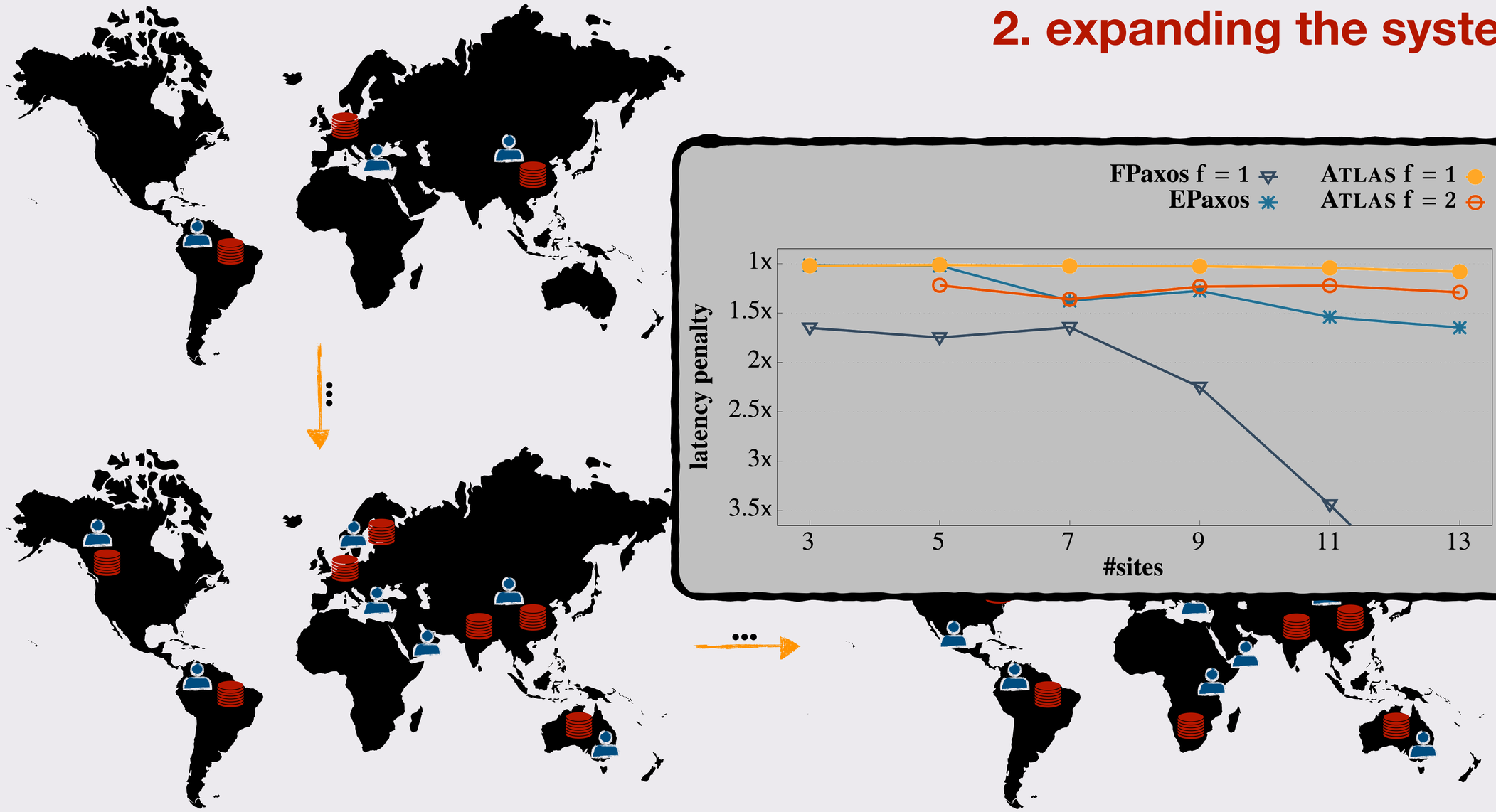
planet-scale evaluation

2. expanding the system



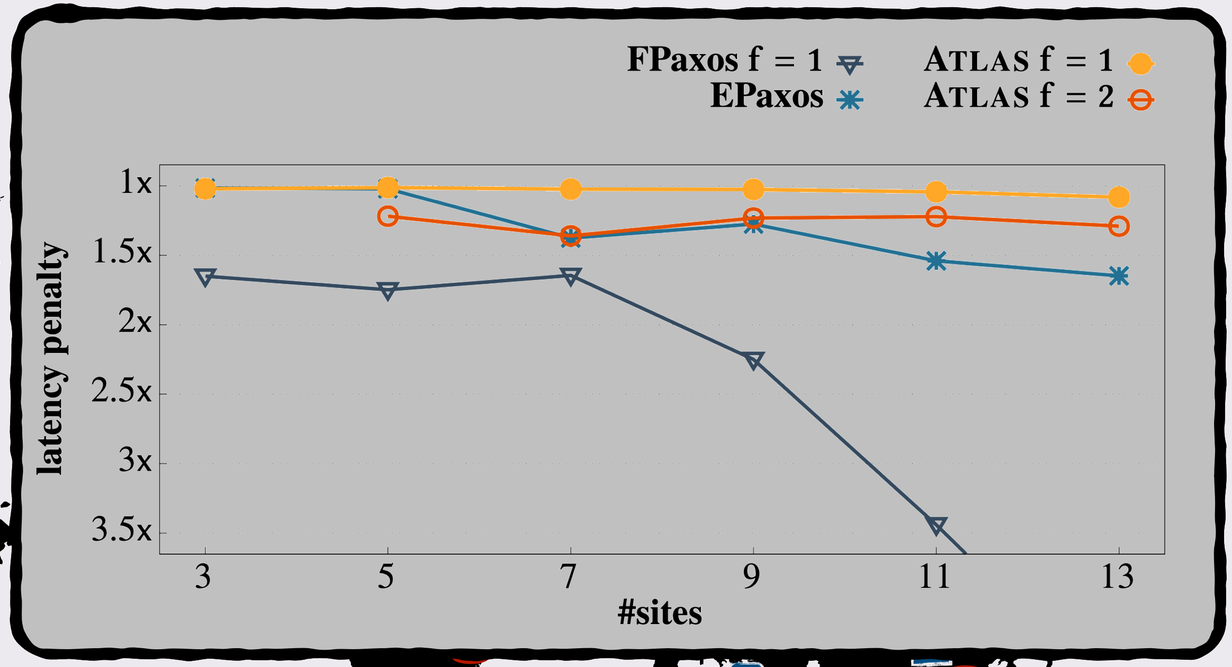
planet-scale evaluation

2. expanding the system



planet-scale evaluation

2. expanding the system



higher is better

atlas copes with growth without degrading performance

more in the paper

- significantly **simpler recovery** than previous protocols
- two **optimizations** that speed up execution
- **evaluation**
  - fast-path likelihood
  - availability under failures
  - YCSB
- **proof of correctness** (arXiv)

summary

- **atlas** exploits the fact that, in **planet-scale** systems, concurrent DC failures are rare
  - first leaderless SMR protocol that allows configuring **n** independently of **f**
  - **small quorums** (for small values of **f**)
  - **flexible fast-path condition** allows a high percentage of commands to be processed in a **single round trip**

state-machine replication for planet-scale systems

Vitor Enes, Carlos Baquero, Tuanir França Rezende,
Alexey Gotsman, Matthieu Perrin, Pierre Sutra

29 Apr. 2020 @ EuroSys'20

vitorenes.org

@vitorenesduarte