Raft Recursive Domains

On how-to replicate between separated Raft clusters

Samo Pogačnik

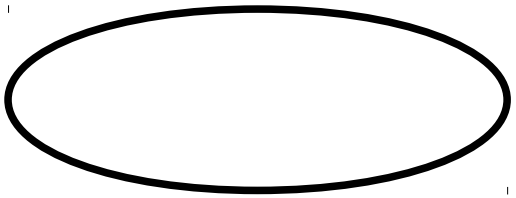
Škofja Loka, 26.11.2016

Legend:

- Leader Cluster Node

- Follower Cluster Node

- Candidate Cluster Node



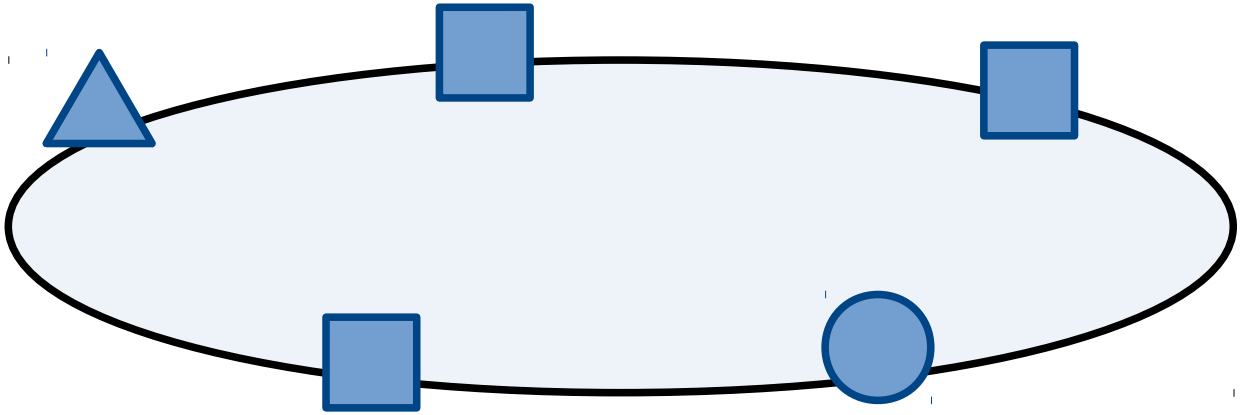
- Cluster Recursive Domain

- Primary Leader Cluster Node

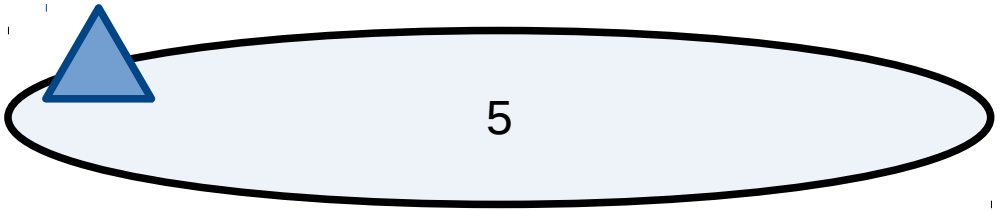
- Secondary Leader Cluster Node

- Secondary/Primary Leader Cluster Node

Sample layouts:

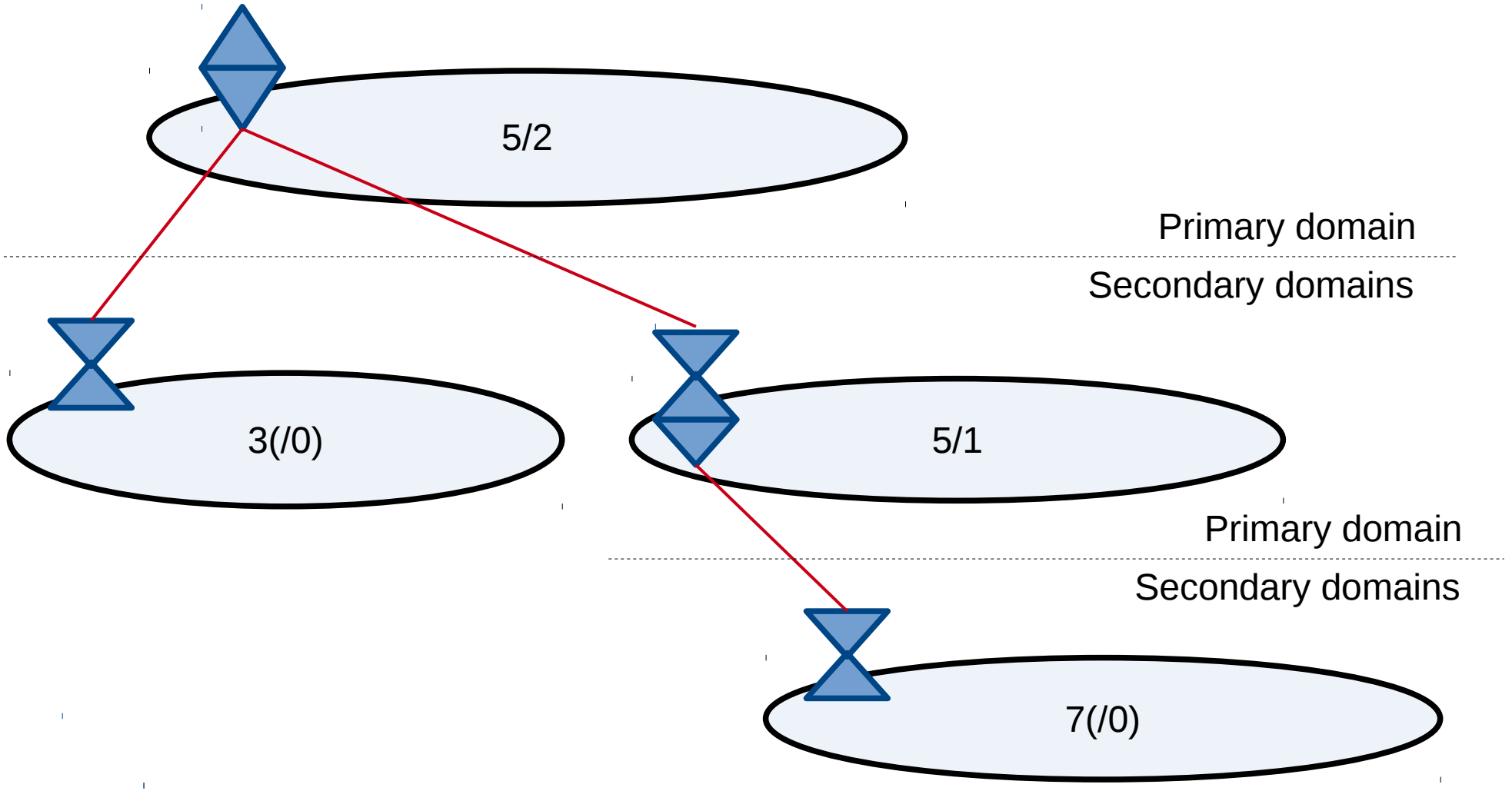


A single 5 node cluster in its current state of node roles.



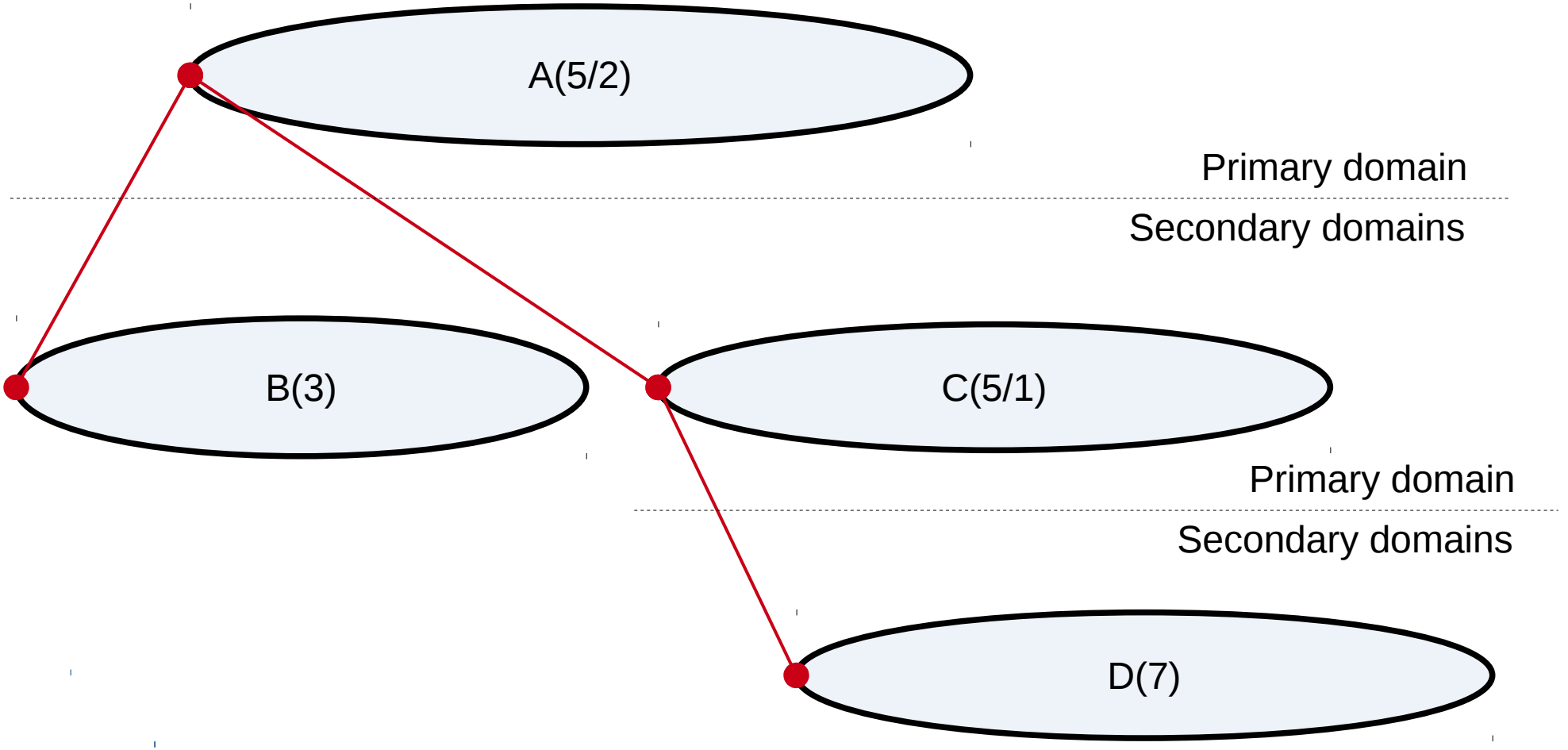
The same single 5 node cluster exposing only its current leader and number of cluster nodes.

Sample layouts:



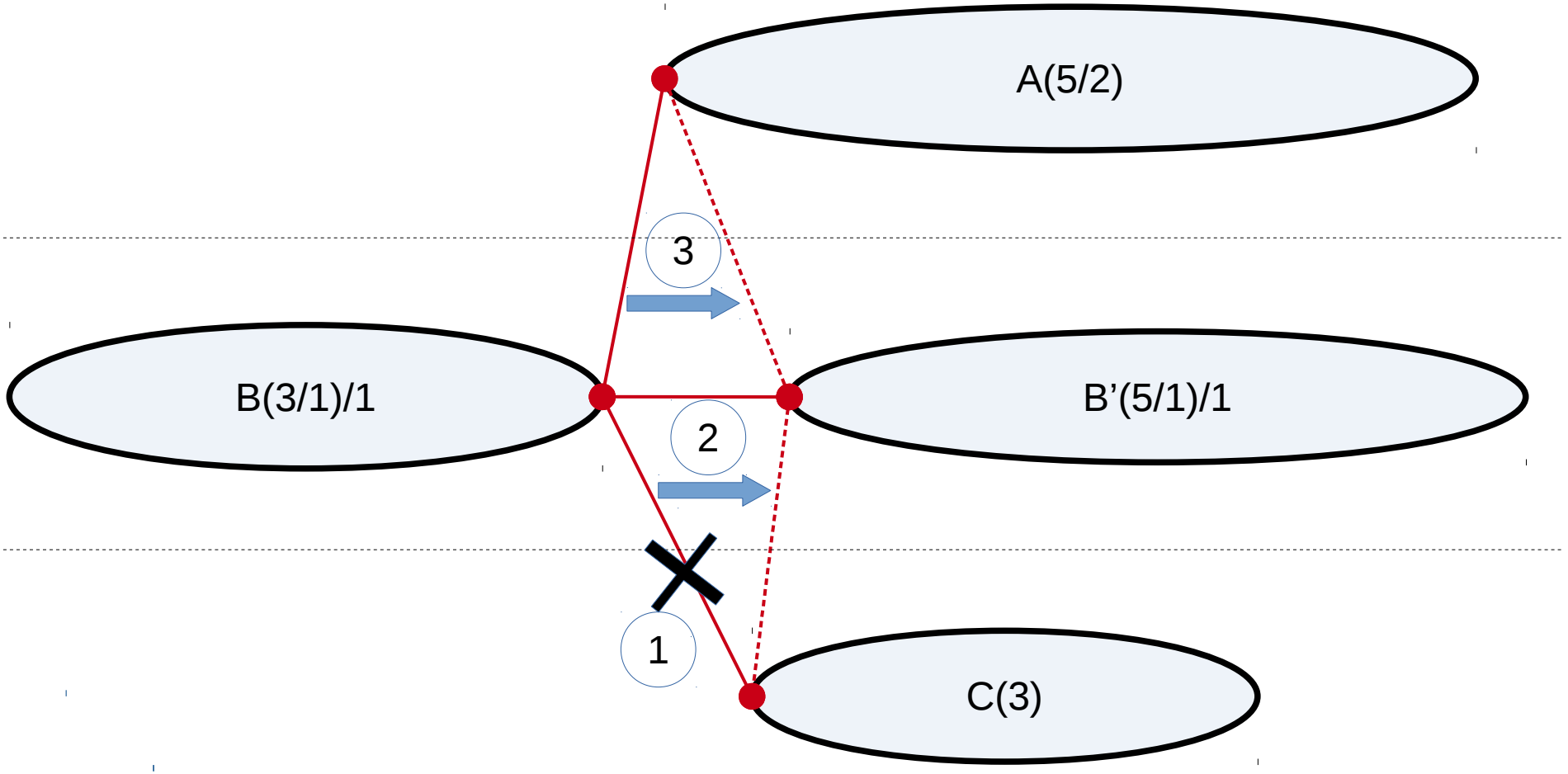
A layout of 5 recursive domains exposing their current leaders, a number of each domain nodes, slash a number of their direct subordinate recursive domains.

Sample layouts – same as previous (no leaders exposed):



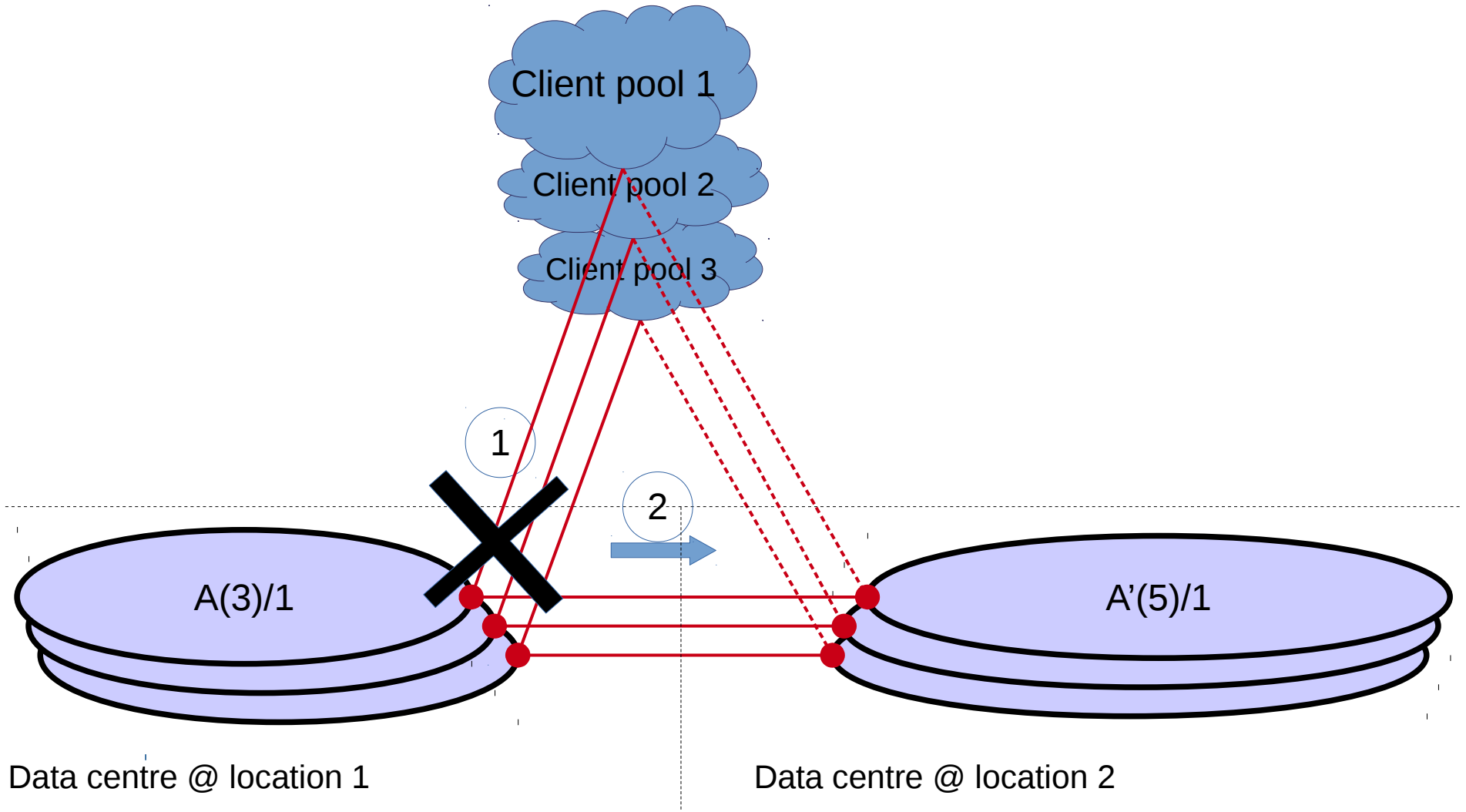
A layout of 5 recursive domains exposing a number of each domain nodes, slash a number of their direct subordinate recursive domains.

Sample layouts (geographically dislocated clusters B and B'):



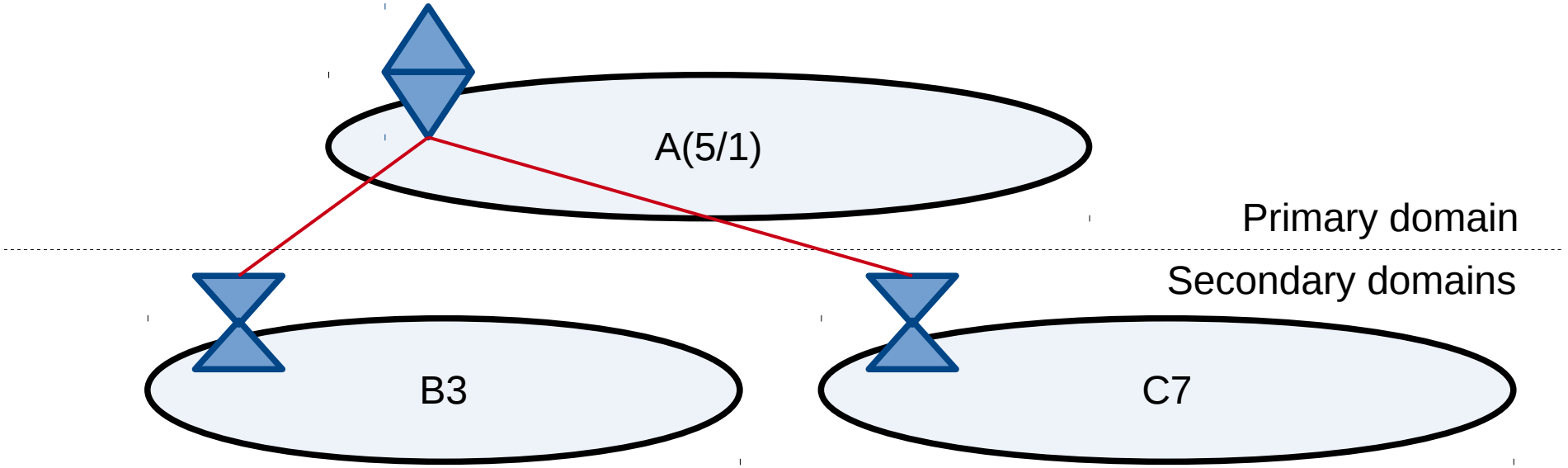
A layout of 4 recursive domains A, B, B' and C. Recursive domains B and B' switch each others Primary/Secondary relation, after connectivity outage between domain B and domain C. What if there is also a domain B''?

Sample layouts (geographically dislocated clusters with clients):



A layout of 3 primary recursive domains A, B and C geographically replicated for disaster recovery scenario (switching primary roles to domains A', B' and C').

Secondary to primary domain connection:



How to include secondary leaders into the primary domain?

1. Awareness of all secondary domains need to be stored within the primary domain!
2. Secondary leaders get elected via normal Raft elections within each secondary domain.
3. Secondary leaders connect to the primary domain's leader via the same mechanisms as clients (i.e. connect to any primary domain node and redirect to their leader) – a unique domain ID separates secondary leaders from clients.

How does a secondary connection affect primary domain?

1. Let's keep things simple and assume static configuration of secondary domains within the primary domain.
2. Also assume asynchronous commits from the primary to all secondary domains.
3. Saying that, any primary-secondary interaction does not count into any quorum (majority) decision (not for primary leader election nor for commit quorum).
4. Only log replication needs to be performed additionally from primary to each secondary leader.
5. A client also does not need to wait for secondary leaders to confirm commit within their secondary domain.

Secondary to primary domain connection (continued):

How to include secondary leaders into the primary domain?

1. A primary domain does not need to know all secondary nodes from each secondary domain.
2. Instead unique domain IDs would be used to recognise each secondary leader, when primary domain nodes are being contacted.
3. Using domain IDs, secondary leaders would be differentiated from clients.

How does a secondary connection affect primary domain?

1. Any secondary leader is always a follower in the primary domain, meaning that secondary domains do not affect primary domain leader elections in any way. Not even during primary domain configuration change, when a secondary domain is added or removed from configuration.
2. What if synchronous commits were required for some secondary domains? In this case, beside the majority of responses from primary followers, additional response from each synchronous secondary domain leader is required too, to consider new entry committed. This kind of operation could be very problematic:
   a) Slow secondary response causes delays in client commits.
   b) Unstable secondary leader causes even more delays in client commits.
   c) A synchronous secondary domain outage blocks all commits.

Secondary leader operation:

H?

1. A.

H?

1. A.

Temporary slide – just elements: