Michael Hauser – DisSys Sheet10 – Team Koeberle, Hauser, Rambacher

a)
Node-to-node communication happens exclusively using the Cluster bus and the Cluster bus protocol: a binary protocol composed of frames of different types and sizes. The Cluster bus binary protocol is not publicly documented since it is not intended for external software devices to talk with Redis Cluster nodes using this protocol.

Also interessting could be that redis cluster exchange ping and pong packets, they have the same struct and owns the same config, the fancy difference could be that the will have different msg type fields. One node is sending to the other node packets, so the receiver gets a trigger to reply if its necessary or if its part of the msg if they should reply.

b)
Every node takes a list of flags associated with other known nodes. There are two flags that are used for failure detection that are called PFAIL and FAIL. PFAIL means Possible failure, and is a non-acknowledged failure type. FAIL means that a node is failing and that this condition was confirmed by a majority of masters within a fixed amount of time.

The Redis Cluster failure detection has a liveness requirement: eventually all the nodes should agree about the state of a given node. There are two cases that can originate from split brain conditions. Either some minority of nodes believe the node is in FAIL state, or a minority of nodes believe the node is not in FAIL state.

The FAIL flag is only used as a trigger to run the safe part of the algorithm for the slave promotion. In theory a slave may act independently and start a slave promotion when its master is not reachable, and wait for the masters to refuse to provide the acknowledgment if the master is actually reachable by the majority.

c)
Redis Cluster uses a concept similar to the Raft algorithm "term". In Redis Cluster the term is called epoch instead, and it is used in order to give incremental versioning to events. When multiple nodes provide conflicting information, it becomes possible for another node to understand which state is the most up to date.

Every time a packet is received from another node, if the epoch of the sender (part of the cluster bus messages header) is greater than the local node epoch, the currentEpoch is updated to the sender epoch.

Because of these semantics, eventually all the nodes will agree to the greates currentEpoch in the cluster. This information is used when the state of the cluster is changed and a node seeks agreement in order to perform some action.

d)
In order to remain available when a subset of master nodes are failing or are not able to communicate with the majority of nodes, Redis Cluster uses a master-slave model where every hash slot has from 1 (the master itself) to N replicas (N-1 additional slaves nodes). Also nice to know, if the master isnt available for a period set by the config, it starts to fail and it wont accept any queries anymore. Due the master slave setup of redis, the limit of storage is set to the minima out of the master slave nodes, and its hard to resize it, if needed.

As we already know from the last exercise sheet – the slave node is assigned to every master, if the master-node fails, its slave node is promoted as the new master, and the cluster will continue to operate correctly.