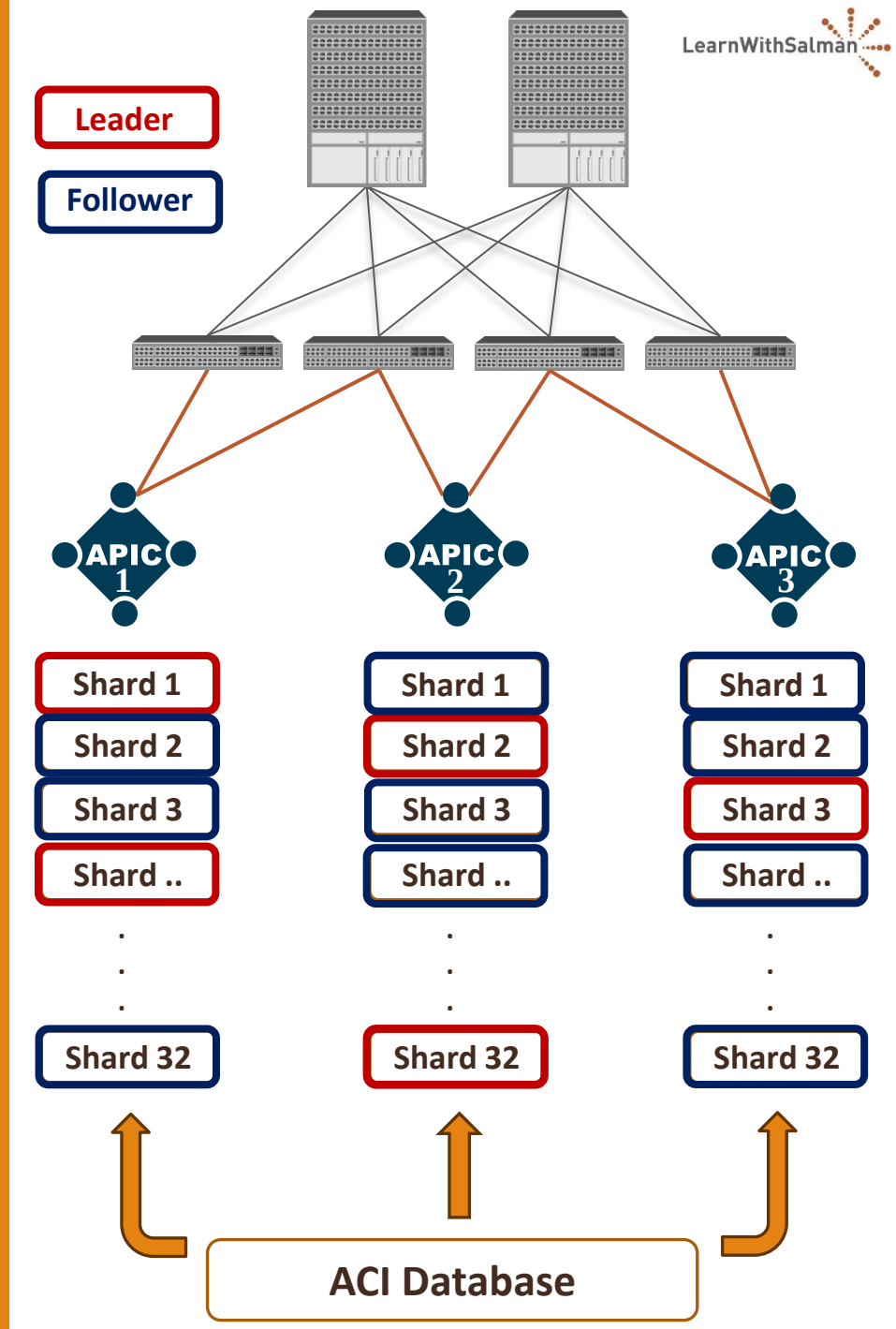
CCIE DATA CENTER

ACI CORE

ACI APIC Clustering & Database Distribution

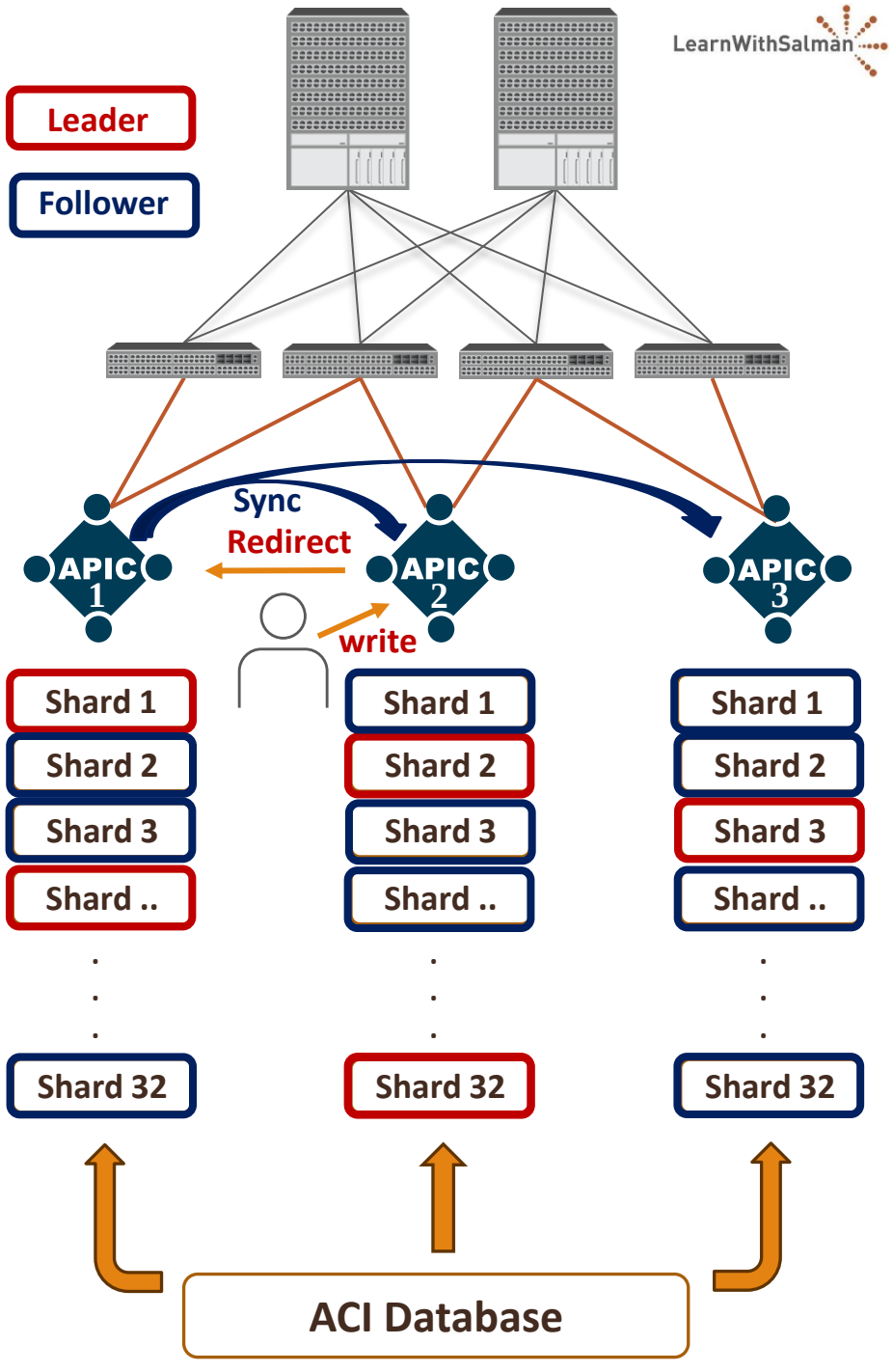
ACI Clustering Overview

- Cisco APIC appliance has two form factors:
  - Medium (M): supports up to 1200 Leaf edge ports.
  - Large (L): supports more than 1200 Leaf edge ports.
- Cisco APICs are deployed as a cluster of servers based on the scalability requirements (up to 5 APICs in a single Pod and up to 7 APICs in Multipod).
  - For high availability, the minimum APIC cluster size should be 3, and larger clusters increase fabric scalability, not high availability.
  - APICs in a cluster discover each other via an LLDP-based discovery process.
- The ACI database is replicated and broken up into smaller database units called **shards**:
  - The ACI database is broken into 32 shards. Each MO is part of a shard.
  - A shard is replicated across 3 APICs regardless of the cluster size. So, every APIC in the 3-APIC cluster will have a copy of the 32 shards. However, this is not the case in larger APIC clusters.
  - For each shard replica, there is one shard leader and two shard followers.
  - One APIC in the cluster will be a shard leader for every individual shard.
  - Shards are evenly distributed among the APICs in the cluster.



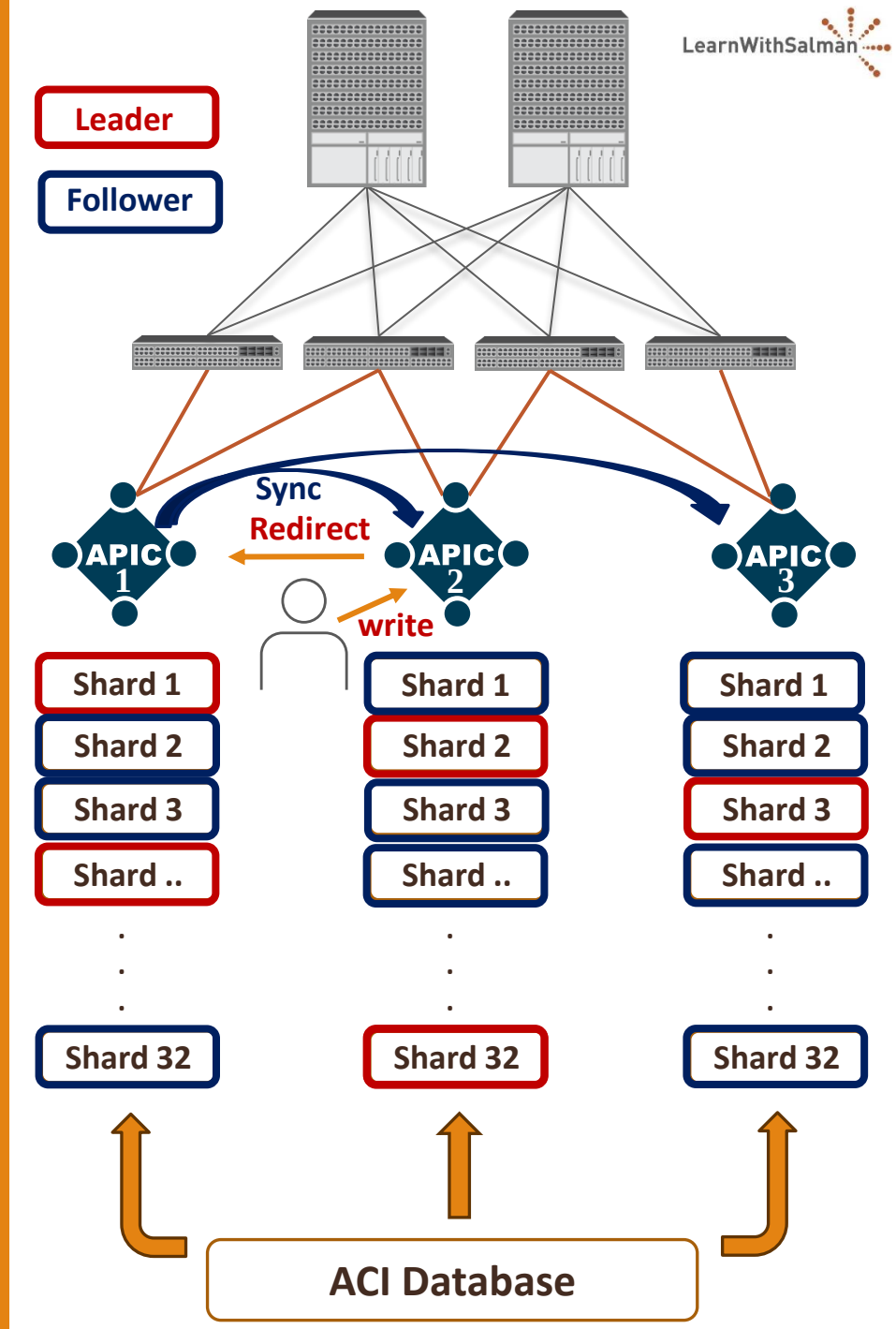
ACI Clustering and Database Distribution

- The 'Shard Leader' APIC is the only APIC that has write access to the shard.
  - Write requests are redirected to the shard leader, who then replicates (sync) the configuration changes to other APICs.
  - So if the admin is connected to APIC2 and tries to write into shard 1, which is led by APIC1, then APIC2 will redirect the write request to APIC1.
  - Read requests can be handled by any APIC (no redirect).



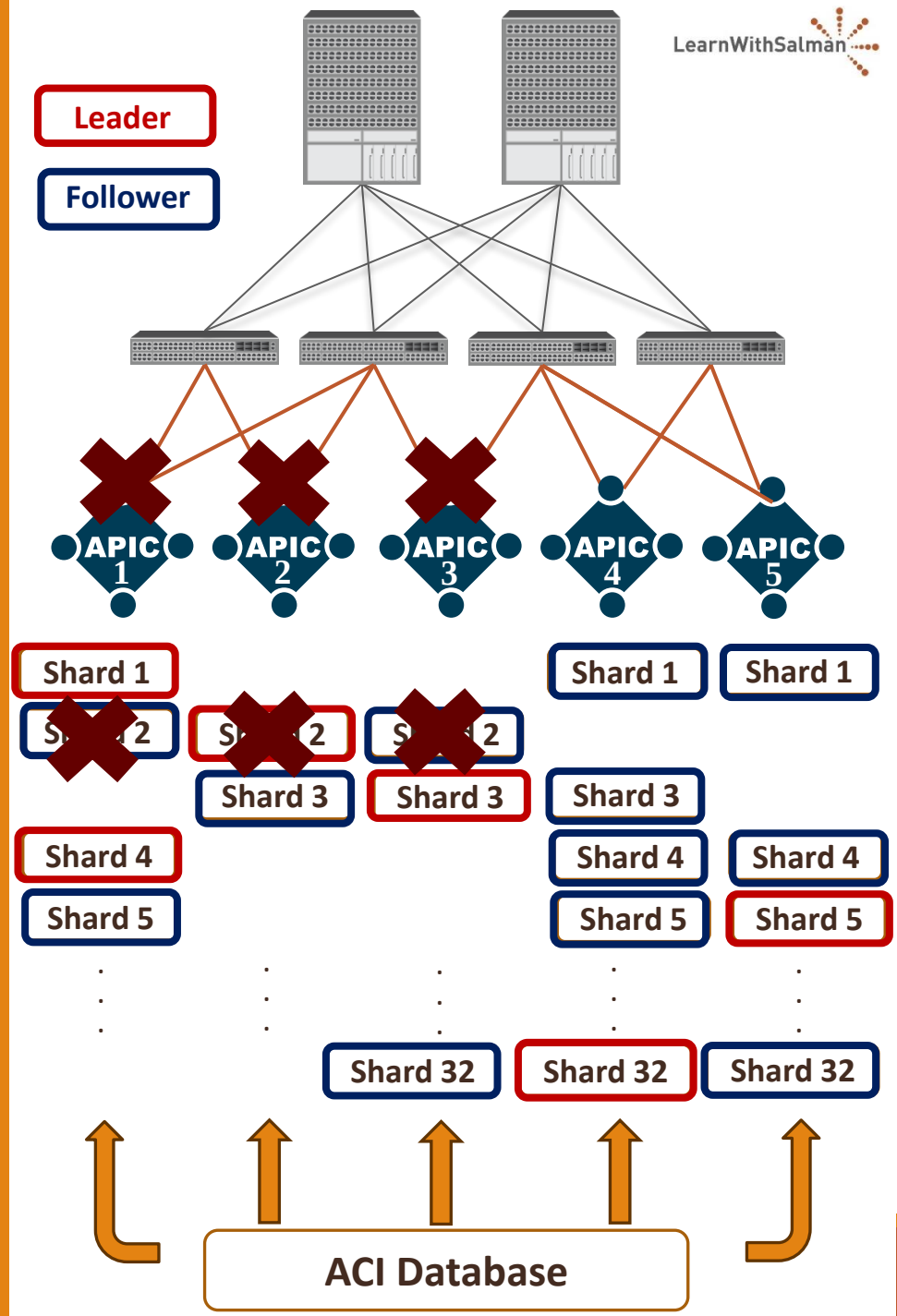
ACI Clustering and Database Distribution

- The 'Shard Leader' APIC is the only APIC that has write access to the shard.
  - Write requests are redirected to the shard leader, who then replicates (sync) the configuration changes to other APICs.
  - So if the admin is connected to APIC2 and tries to write into shard 1, which is led by APIC1, then APIC2 will redirect the write request to APIC1.
  - Read requests can be handled by any APIC (no redirect).
- When the cluster size exceeds 3 APICs, we still have only three replicas for each shard.



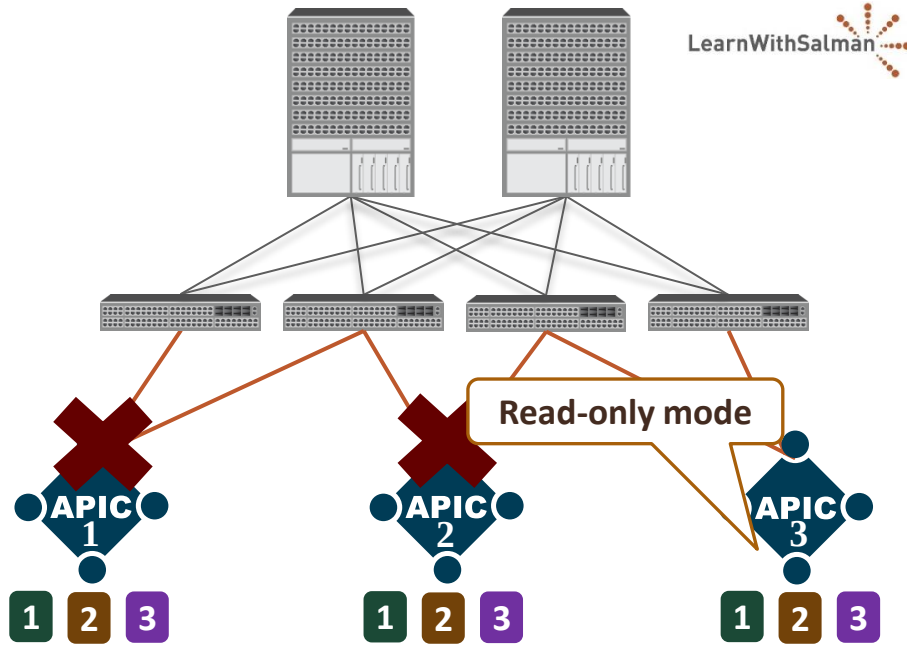
ACI Clustering and Database Distribution

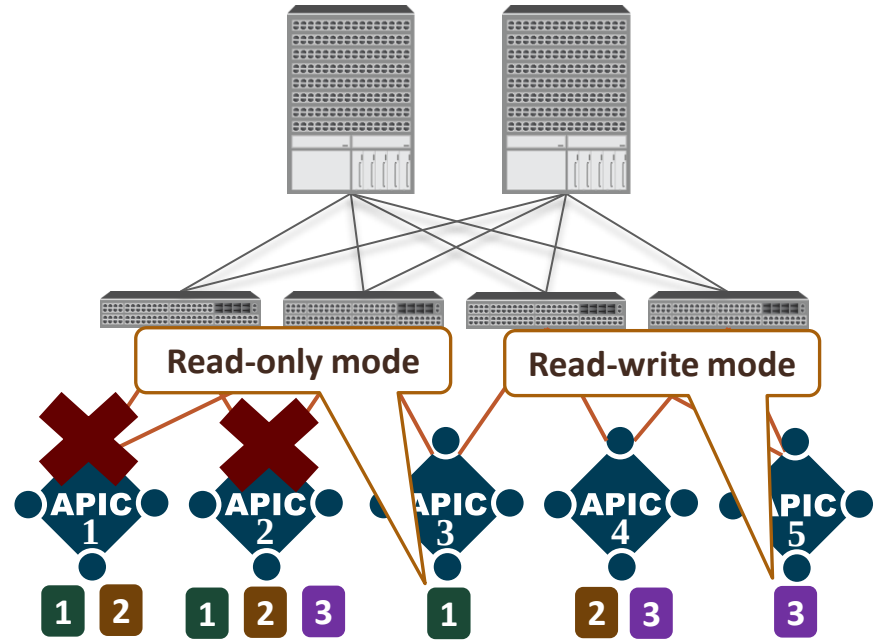
- The 'Shard Leader' APIC is the only APIC that has write access to the shard.
  - Write requests are redirected to the shard leader, who then replicates (sync) the configuration changes to other APICs.
  - So if the admin is connected to APIC2 and tries to write into shard 1, which is led by APIC1, then APIC2 will redirect the write request to APIC1.
  - Read requests can be handled by any APIC (no redirect).
- When the cluster size exceeds 3 APICs, we still have only three replicas for each shard.
  - In the case of the 5-APIC cluster, for every Shard, we have 2 APICs that don't have a copy of it.
  - If any three APICs go down, we will lose a shard (all three replicas), which means part of the database is completely lost.
  - So, having more APICs doesn't mean high availability but more scalability.
- ACI database distribution benefits:
  - Shards provide faster reads and writes operations.
  - Load balancing and more scalable since write operations will get distributed across APICs (Leaders).



APIC Clustering Failure Considerations

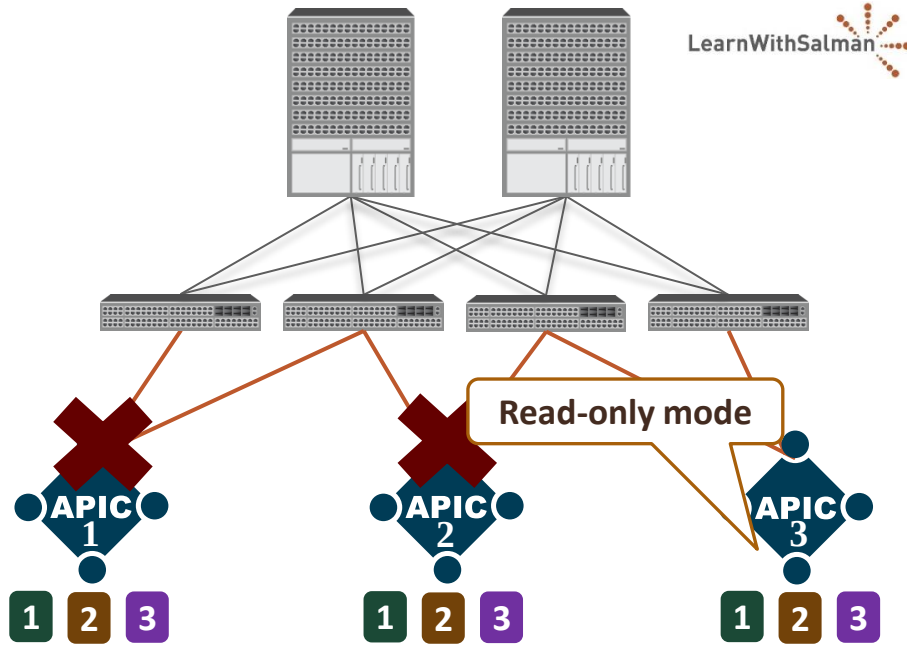
- When APIC failure happens, all shards led by the faulty APIC will be evenly led by the active APICs.
  - If APIC 3 goes down, all shards led by APIC 3 will be led sequentially between APIC 1 and APIC 2.
- ACI allows ready-only access to a given shard when only one replica remains for that shard (**minority state**).
  - Losing two APICs in a 3-APIC cluster causes all shards to have only one replica, therefore, ready-only access to all shards (the whole Database).
  - Losing two APICs in a 5-APIC cluster causes some shards to have only one replica, therefore, inconsistent behavior across shards (some read-only, some read-write access).
- A failure of two APICs will cause some shards to be in the minority state, and losing three APICs will cause some/all shards to be lost.
- To avoid merging issues, no write request is allowed during the minority state.
  - No configuration change.
  - No vCenter update is handled.
  - No updates from switches are handled.

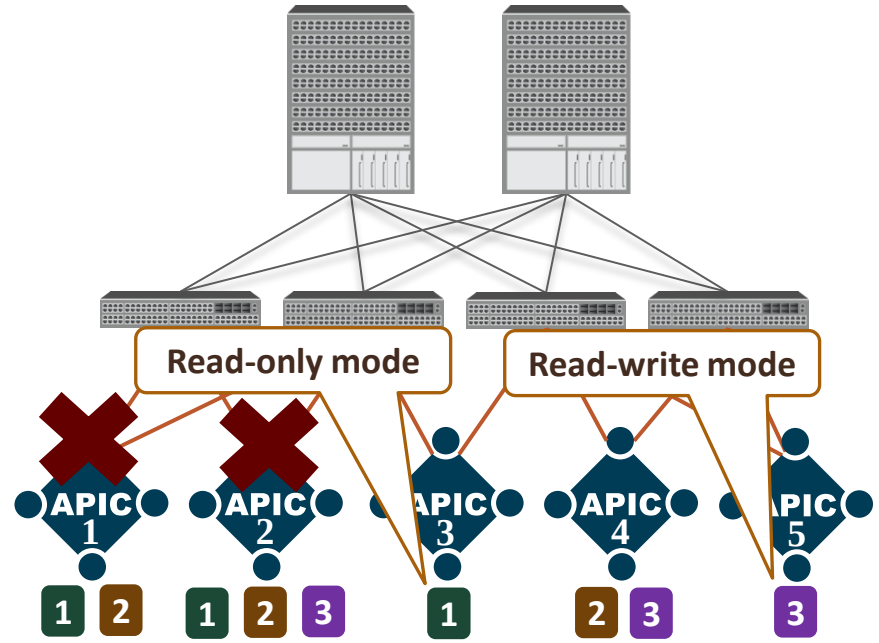




APIC Clustering Failure Considerations (cont.)

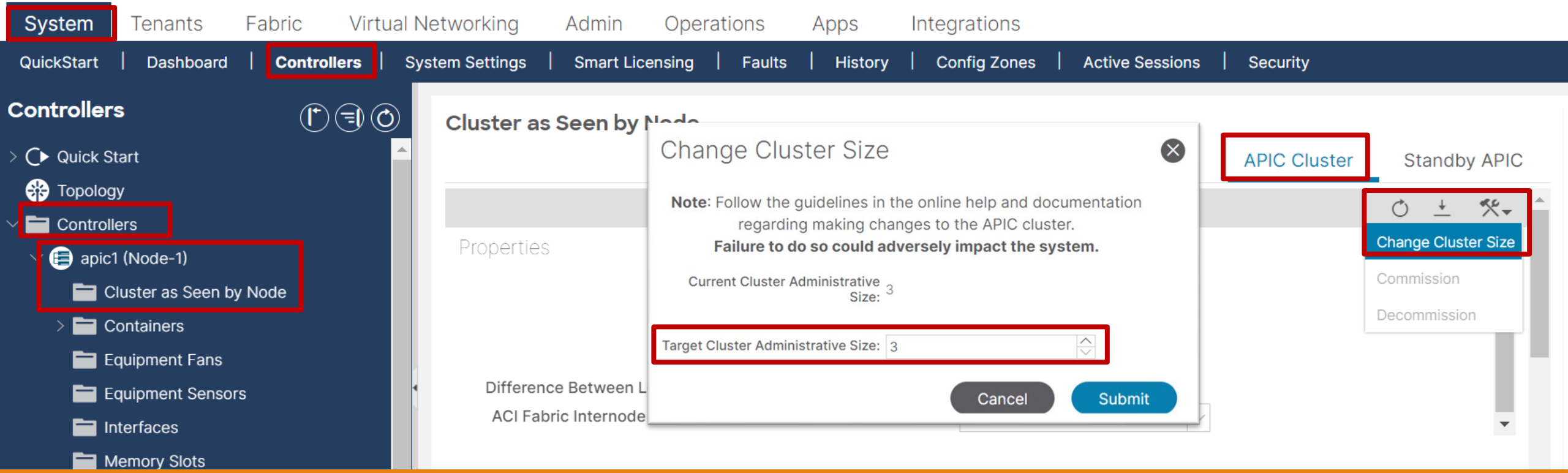
- Why do we need an odd number of active APICs in the cluster?
  - Since shard leaders are distributed in three APICs, we need the majority of replicas to agree on the shard leader in order to have write access during a single failure.
- Why do we need a minimum of three APICs?
  - One APIC: obviously, no fault tolerance. Data lost for a single failure.
  - Two APICs: Write unavailability with a single failure.
  - Three APICs: If one APIC is lost, the other two can elect a new leader and continue writing. If two APICs are lost, we go to minority mode.
- What if all APICs go down?
  - Traffic forwarding continues for new and existing sessions.
  - New VMM endpoint attachment and vMotion may or may not work depending on the configuration options.
  - If you have a configuration snapshot, you can recover the fabric (Fabric ID recovery) with the help of Cisco TAC.
- What if all APICs have not yet been installed (Discovered)?
  - When starting a new fabric, the shards are not in a minority state because the cluster has not yet been fully fit. (only APIC1 with one replica per shard).
  - APIC1 will create new replicas when a new APIC is added to the cluster.





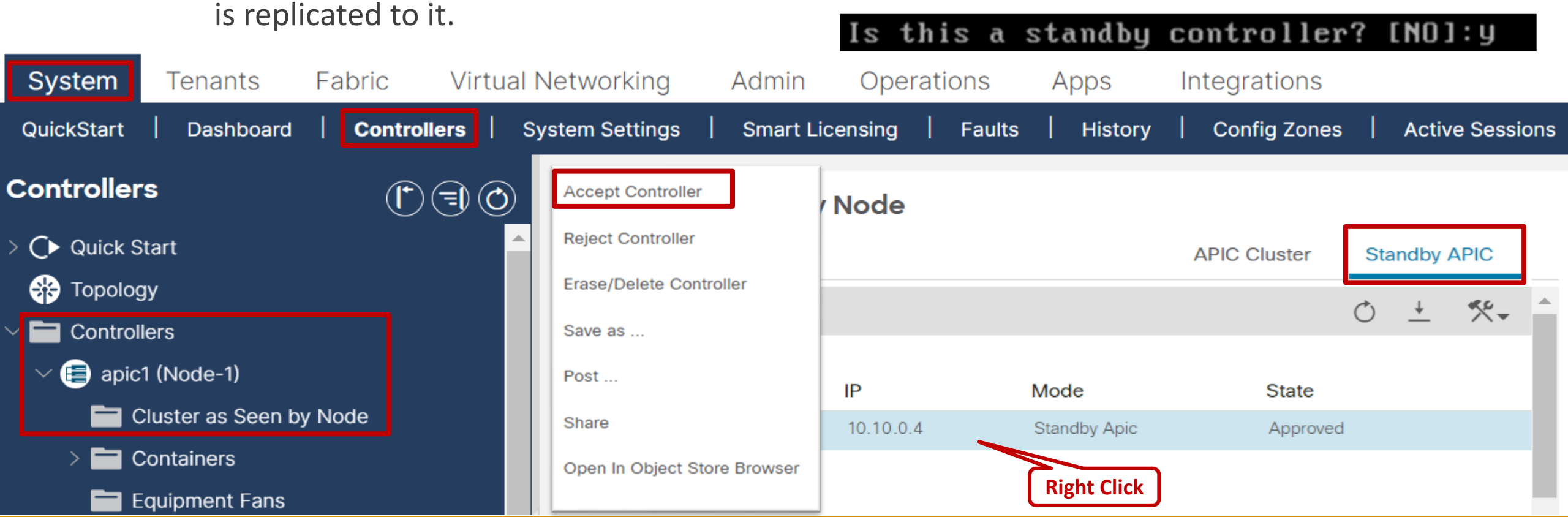
ACI APIC Cluster Resize

- The Cisco APIC can expand and shrink a cluster by defining a target cluster size.
  - When a Cisco APIC cluster is expanded, some shard replicas shut down on the old APICs and start on the new APICs to help distribute evenly across all APICs in the cluster.
  - When removing an APIC from the cluster, we must remove the appliance at the end.



Standby APIC

- The standby Cisco APIC is a controller that you can keep as a spare, ready to replace any active APIC in a cluster in one click.
  - This controller does not participate in policy configurations or fabric management. So, no data is replicated to it.



Who Is the Shard Leader For a Specific Shard ID?

```
"acidiag rvread" shows replica which are not heathly
"acidiag rvread <srv> <shard> <replica>" to see the state of the replicas
```

```
apic1# acidiag rvread 6 1
(6,1,1)   st:6 lm(t):1(2023-11-28T07:15:50.759+00:00) le: reSt:LEADER voGr:0 cuTerm:0x1be lCoTe:0x1bd
lCoIn:0x80000023d7a825a veFi                    En:0x6 lm(t):1(2023-11-28T07:15:50.244+00:00) stMmt:1 lm(t):0(zeroTime)
ReTx:0 lm(t):0(zeroTime) lastUpdt 2024-01-18T20:28:39.444+00:00
(6,1,2)   st:6 lm(t):4(2023-                    :28.804+00:00) le: reSt:FOLLOWER voGr:0 cuTerm:0x1be lCoTe:0x1bd
lCoIn:0x80000023d7a825a veFi                    En:0xa lm(t):4(2023-11-28T07:15:50.487+00:00) stMmt:1 lm(t):0(zeroTime)
ReTx:0 lm(t):0(zeroTime) lastUpdt 2024-01-18T20:28:39.444+00:00
(6,1,3)   st:6 lm(t):5(2023-11-17T09:01:37.429+00:00) le: reSt:FOLLOWER voGr:0 cuTerm:0x1be lCoTe:0x1bd
lCoIn:0x80000023d7a825a veFiSt:0xa veFiEn:0xa lm(t):5(2023-11-28T07:15:50.501+00:00) stMmt:1 lm(t):0(zeroTime)
ReTx:0                                          astUpdt 2024-01-18T20:28:39.444+00:00

apic1# acidiag rvreadle
Optimal leader for all shards
----------------------------------------
clusterTime=<diff=-2924 common=2024-01-20T20:22:10.634+00:00 local=2024-01-20T20:22:13.558+00:00
pF=<displForm=0 offsSt=0 offsVlu=0 lm(t):3(2023-12-10T04:28:31.525+00:00)
```

APIC ID

State, 6=UP

Service ID, shard, replica

Healthy

Thanks for watching!


LearnWithSalman