

# CCDUG

2024 Central Canada Db2 Tech  
Conference

**Intro to Pacemaker, the cluster  
manager of the future!**

**Tharmiga Loganathan**

*IBM Canada Ltd.*

Platform: Db2 LUW

## Please note

- IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM's sole discretion.
- Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.
- The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.
- The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
- Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

# NOTICE AND DISCLAIMER

- © 2024 International Business Machines Corporation. No part of this document may be reproduced or transmitted in any form without written permission from IBM.
- **U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.**
- Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed “as is” without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.
- IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply.”
- **Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.**
- Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.
- References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.
- Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.
- It is the customer’s responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer’s business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

**01**

**INTRODUCTION**

**02**

**PACEMAKER VS. TSA**

**03**

**ARCHITECTURE**

**04**

**FAILOVER BEHAVIOUR**

**05**

**UP AND RUNNING**

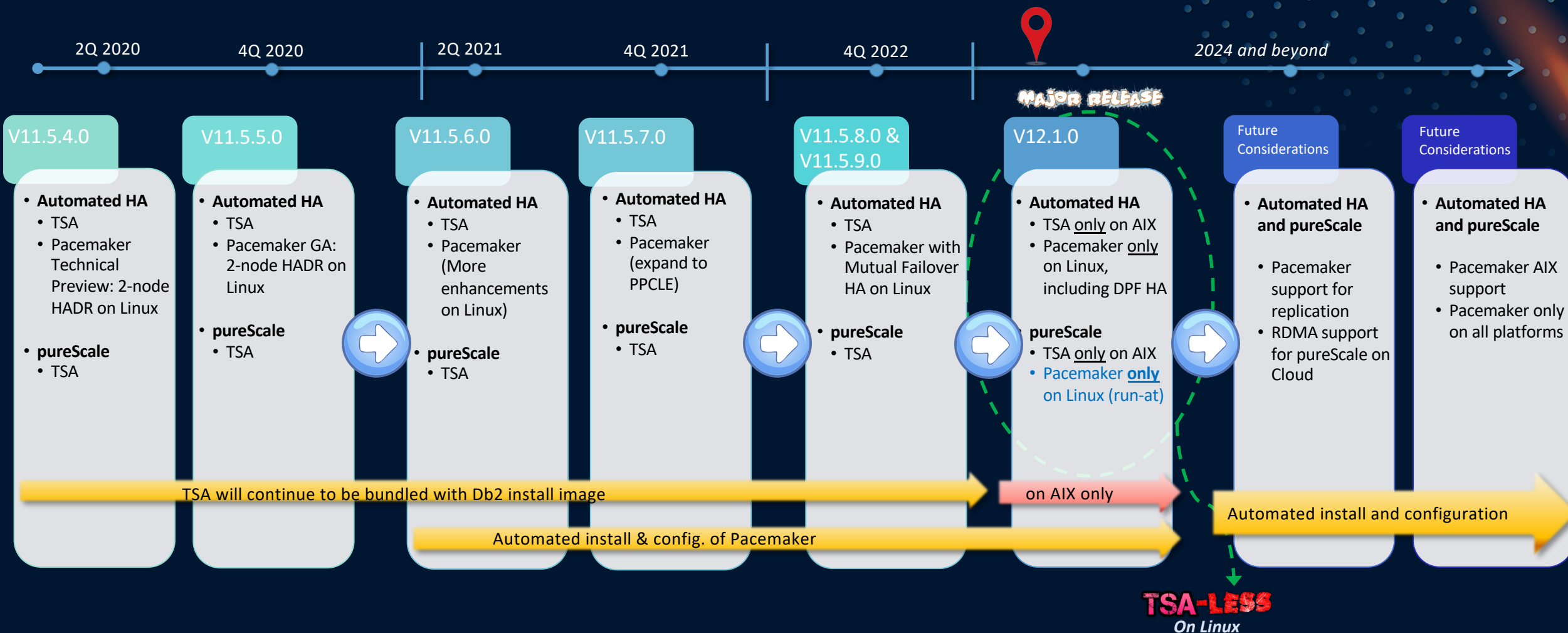
**06**

**DEMO**



# AGENDA

# Db2 Pacemaker Journey



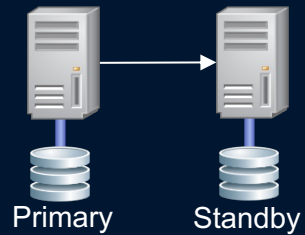
Note: Roadmap subjected to change

# Db2 Cloud-Ready *Integrated* HA Topologies with Pacemaker – 10,000' view

Today's Presentation

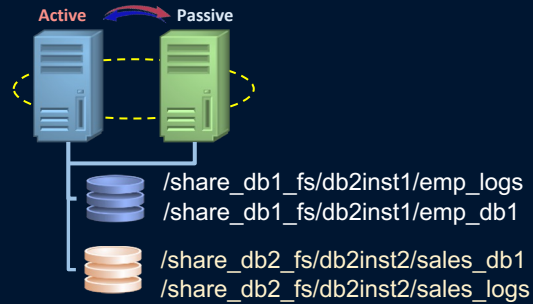
## V11.5.5.0

Single DB Partition (EE)  
with automated HADR



## V11.5.8.0

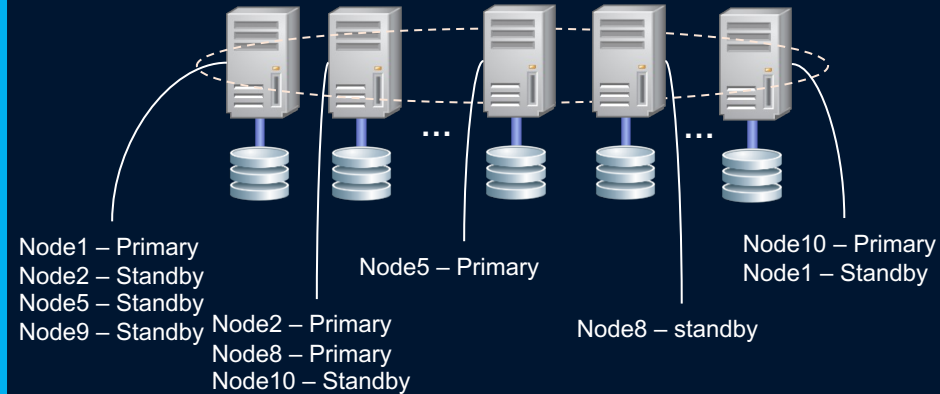
Mutual Failover (MF) (a.k.a. Active/Passive)  
automated HA with shared storage



Today's Demo

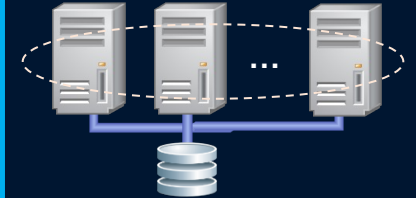
## V12.1.0

Database Partitioning Feature (DPF)  
with automated HA (same site)



## V12.1.0

pureScale  
Online 24x7x365 with  
automatic failover



Our vision with Pacemaker ...

## Db2 Integrated Cluster Manager of Choice

PAST PRESENT FUTURE



TSA  
RSCT

Corosync  
The Corosync Cluster Engine  
Pacemaker

Corosync  
The Corosync Cluster Engine  
Pacemaker



v11.5.8.0

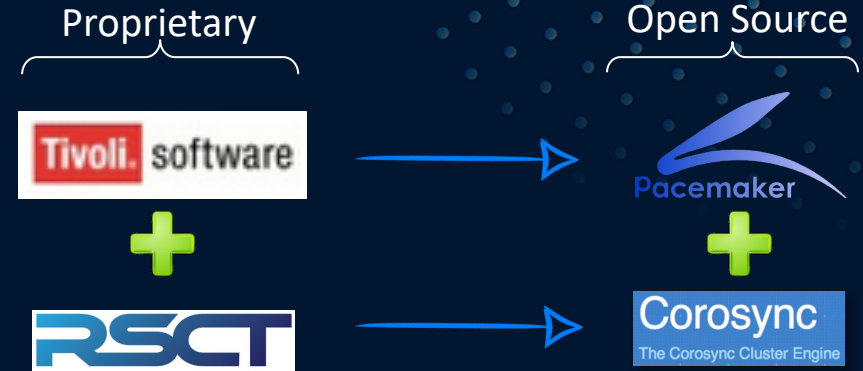
- Announcement of Deprecation of TSA Support on Linux
  - Target: TSA will no longer be bundled with Db2 on Linux in v12.1.0

# Why Pacemaker over TSA ?

- 18+ years in industry as HA resource cluster manager
- Included by RHEL and SuSE as paid add-on HA package
- Open source: allow for future port to AIX
- Align with IBM Open-Source Strategy

## Key driver for change

- Surge of requests for cloud support
- Lack of flexibility with TSA
- Need single solution for all OSES, architectures, form factors



B E N E F I T S	<p>Cloud Ready</p>	<p>PERFORMANCE</p> <p>Faster Failover Recovery</p>	<p>Lower \$\$\$</p>
	<p>First contribution to community in V11.5.6.0 !!!!</p> <p>STACK MODERNIZATION open source</p>	<p>One Team Just Db2</p> <p>Technical Support</p>	<p>Simpler Cluster Software Architecture</p> <p>Simplified PD</p>



# Recovery Performance compared with TSA



- Dual Reboot - ~45%
- Standby Reboot - ~28%
- Software Failure Primary - ~33%
- Software Failure Standby - ~31%
- User initiated TAKEOVER - ~24%



## Mutual Failover

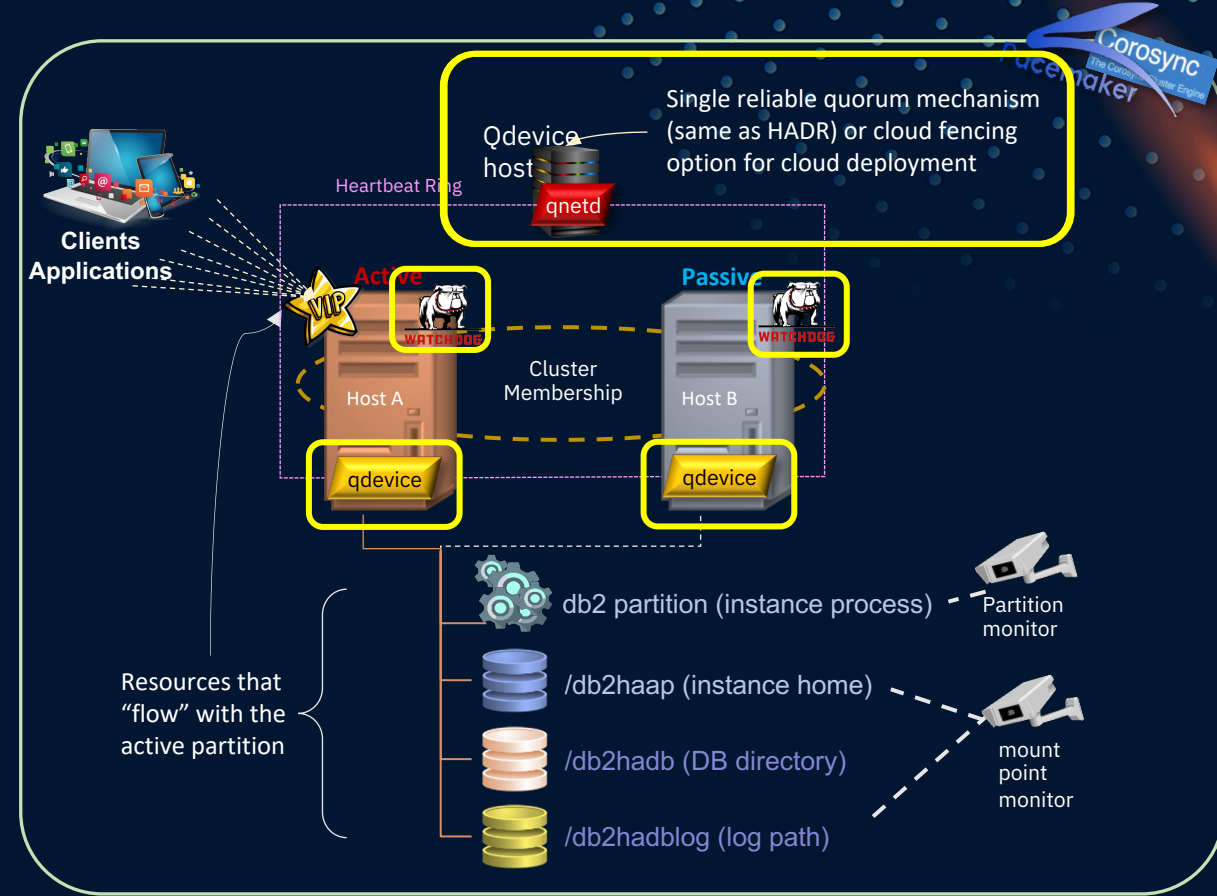
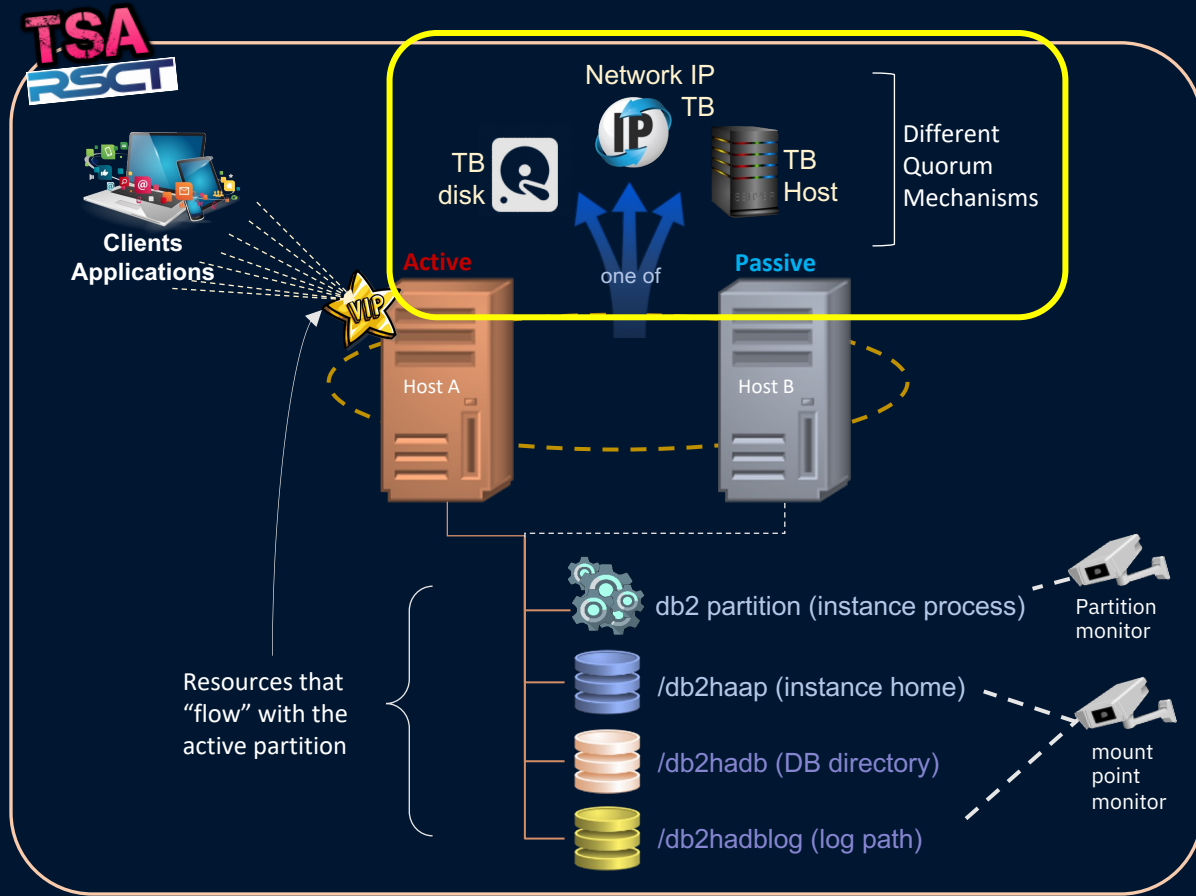
- Reboot - **~155% !!!**
- Software Failure - 29%
- User initiated TAKEOVER - ~50  
seconds in Pacemaker, NOT  
implemented in TSA

Performance result measured from start of test scenario to transaction resumes



*Note: More improvements possible with more experimentation with various config parameters.*

# Old Vs New: *Topology & Components*



TB disk: requires SCSI 2/3 (not cloud-friendly), Network IP: not reliable, TB host: heavy handed

Rely on RSCT Critical Resource Protection to reboot when a resource failed

Engine is integrated, but not every Db2 utility is cluster-aware

No

- Quorum
- Fencing
- Engine Integration
- Cloud-Readiness
- Cluster Management

Qdevice has lightweight non-H/W requirements, reliable, cloud-ready, multi-clusters, cross-arch.

Leverage OS Software Watchdog to reboot host when a fencing action is required

Engine and utilities are cluster-aware such as db2relocatedb

Yes

db2cm – option-based command line utility

# Old Vs New: *Resource Model*



```

Online IBM.Equivalency:db2_private_network_0
|- Online IBM.NetworkInterface:eth0:host1
'- Online IBM.NetworkInterface:eth0:host2

Online IBM.ResourceGroup:db2_db2inst1_0-rg Nominal=Online

|- Online IBM.Application:db2_db2inst1_0-rs
|- Online IBM.Application:db2_db2inst1_0-rs:host1
'- Offline IBM.Application:db2_db2inst1_0-rs:host2

'- Online IBM.Application:db2mnt-db2haap-rs
|- Online IBM.Application:db2mnt-db2haap-rs:host1
'- Offline IBM.Application:db2mnt-db2haap-rs:host2

'- Online IBM.Application:db2mnt-db2hadb-rs
|- Online IBM.Application:db2mnt-db2hadb-rs:host1
'- Offline IBM.Application:db2mnt-db2hadb-rs:host2

'- Online IBM.Application:db2mnt-db2hadbblog-rs
|- Online IBM.Application:db2mnt-db2hadbblog-rs:host1
'- Offline IBM.Application:db2mnt-db2hadbblog-rs:host2

'- Online IBM.ServiceIP:db2ip_9_26_124_190-rs
|- Online IBM.ServiceIP:db2ip_9_26_124_190-rs:host01
'- Offline IBM.ServiceIP:db2ip_9_26_124_190-rs:host02
    
```

```

Full List of Resources:
* db2_host1_eth0 (ocf::heartbeat:db2ethmon): Started host1
* db2_host2_eth0 (ocf::heartbeat:db2ethmon): Started host2

* db2_db2inst1_0 (ocf::heartbeat:db2partition): Started host1

* db2_db2inst1_0-mnt_db2haap (ocf::heartbeat:db2fs): Started host1
* db2_db2inst1_0-mnt_db2hadb (ocf::heartbeat:db2fs): Started host1
* db2_db2inst1_0-mnt_db2hadbblog (ocf::heartbeat:db2fs): Started host1

* db2_db2inst1_0-VIP (ocf::heartbeat:IPaddr2): Started host1

colocation co-db2_db2inst1_0-with-db2_db2inst1_0-mnt_db2haap inf:
db2_db2inst1_0:Started db2_db2inst1_0-mnt_db2haap:Started
order or-db2_db2inst1_0-mnt_db2haap-then-db2_db2inst1_0 Mandatory:
db2_db2inst1_0-mnt_db2haap db2_db2inst1_0

colocation co-db2_db2inst1_0-with-db2_db2inst1_0-mnt_db2hadb inf:
db2_db2inst1_0:Started db2_db2inst1_0-mnt_db2hadb:Started
order or-db2_db2inst1_0-mnt_db2hadb-then-db2_db2inst1_0 Mandatory:
db2_db2inst1_0-mnt_db2hadb db2_db2inst1_0

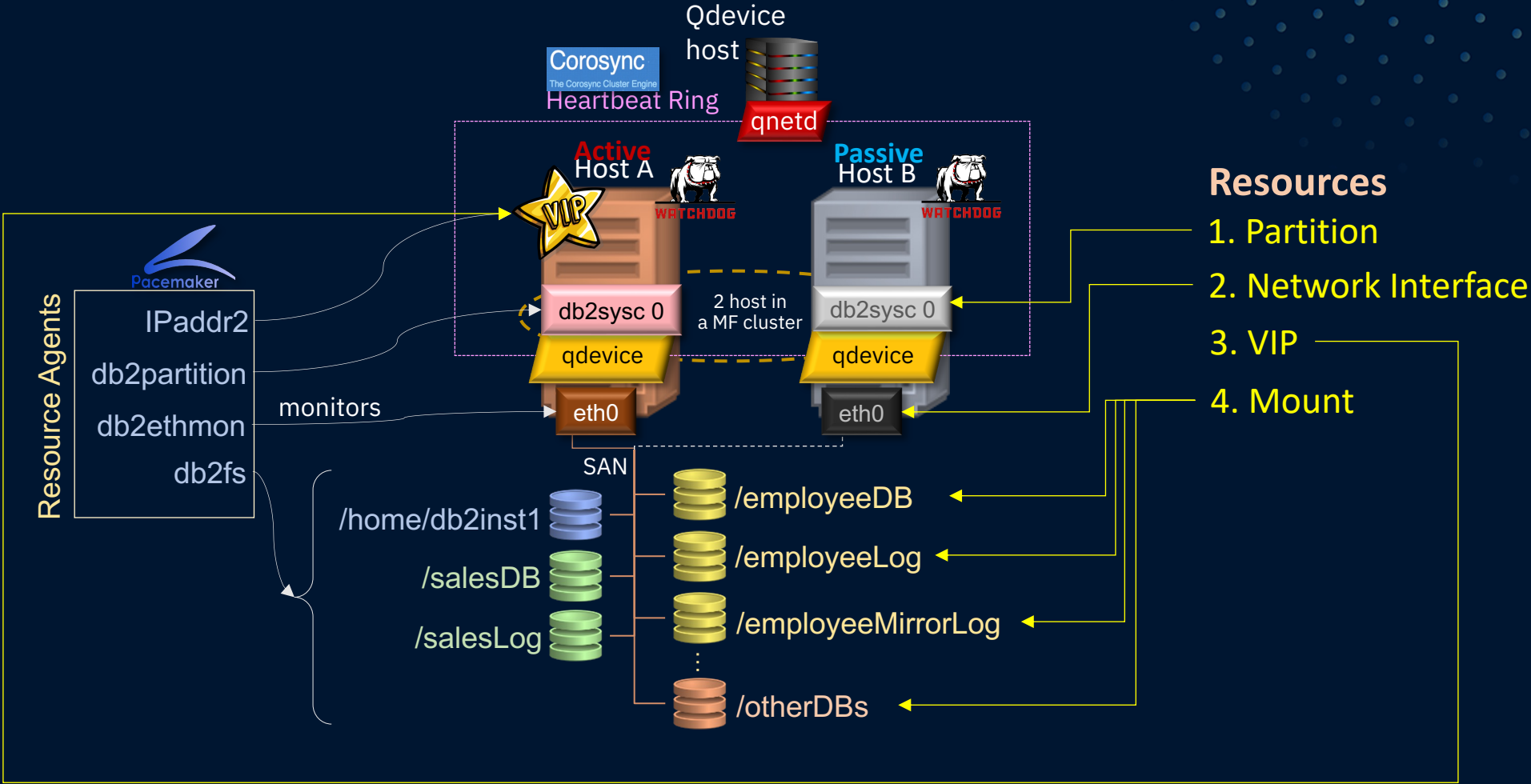
colocation co-db2_db2inst1_0-with-db2_db2inst1_0-mnt_db2hadbblog inf:
db2_db2inst1_0:Started db2_db2inst1_0-mnt_db2hadbblog:Started
order or-db2_db2inst1_0-mnt_db2hadbblog-then-db2_db2inst1_0 Mandatory:
db2_db2inst1_0-mnt_db2hadbblog db2_db2inst1_0

colocation co-db2_db2inst1_0-VIP-with-db2_db2inst1_0 inf:
db2_db2inst1_0-VIP:Started db2_db2inst1_0:Started
order or-db2_db2inst1_0-then-db2_db2inst1_0-VIP Mandatory:
db2_db2inst1_0:start db2_db2inst1_0-VIP:start
    
```

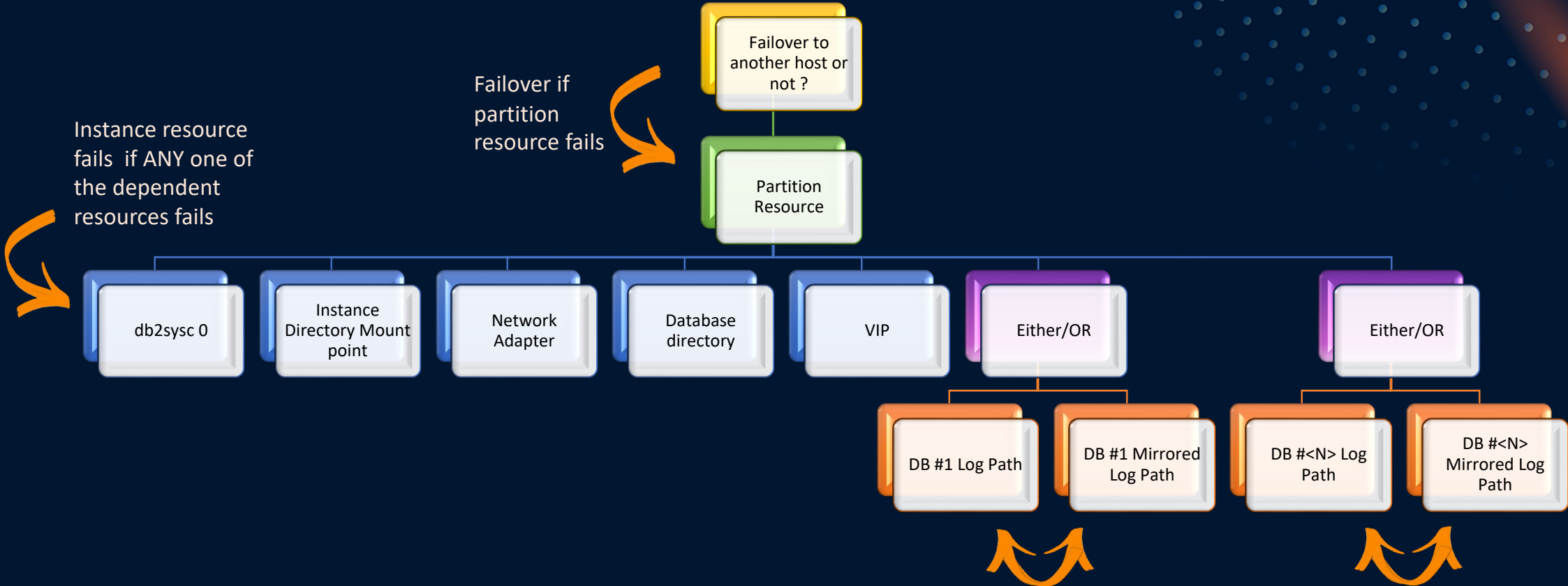


- Behaviours:**
- Every non-network resources tie to the partition resource
  - Every resource belonging to the partition is active on one and only one host at a time.
- 11• No Db level monitoring.

# Architecture: Overview of Resource Models Components



# Architecture: Resources Dependencies



Instance resource fails if ANY one of the dependent resources fails

Failover if partition resource fails



**Note:** A local restart of the resource is always the first action by Pacemaker before a failover will be attempted.

Overall Log Path is considered online if one of the Log Path or Mirrored Log Path mount point is online

Overall Log Path is considered online if one of the Log Path or Mirrored Log Path mount point is online

## Architecture: Resource Model: *Resource Agents*

- A set of shell scripts developed and supported by Db2 to perform actions on the resources defined in the resource model.
- A total of four resource agents installed in `/usr/lib/ocf/resource.d/heartbeat/`:

### *IPaddr2*

- Monitor, start, and stop the VIP resource

### *db2partition*

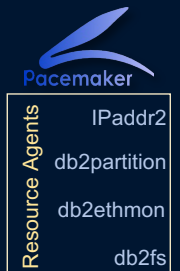
- Monitor, start, and stop a Db2 partition. Verifies the state of the partition (online/offline) and takes the required action to bring up the partition if needed. This is at instance level.

### *db2ethmon*

- Monitor, start, and stop the network adapter resource.
- Same agent as for HADR

### *db2fs*

- Monitor, start, and stop individual mount points.



# Architecture: Inferring the cluster topology from "db2cm -list" - *Resource Model*

```
$ ./db2cm -list
Cluster Status
```

HA configuration = Mutual Failover

```
Domain information:
Domain name          = db2domain
Pacemaker version    = 2.1.2-4.db2pcmk.el8
Corosync version     = 3.1.6
Current domain leader = lcars-srv-1
Number of nodes      = 2
Number of resources  = 5
```

Node information: Cluster node list

Name	State
lcars-srv-1	Online
lcars-srv-2	Online

Resource Information:

Network adapters

```
Resource Name = db2_lcars-srv-1_eth0
State         = Online
Managed      = true
Resource Type = Network Interface
Node         = lcars-srv-1
Interface Name = eth0

Resource Name = db2_lcars-srv-2_eth0
State         = Online
Managed      = true
Resource Type = Network Interface
Node         = lcars-srv-2
Interface Name = eth0
```

Active Partition

```
Resource Name = db2_regress1_0
State         = Online
Managed      = true
Resource Type = Partition
Instance     = regress1
Partition    = 0
Current Host  = lcars-srv-1
```

```
Resource Name = db2_regress1_0-instmnt_db2hamf
State         = Online
Managed      = true
Resource Type = File System
Device       = "/dev/disk/by-
uuid/f3983a2b-fbf6-4b10-8032-0cecc2e044fb"
Mount Point  = "/db2hamf"
File System Type = ext3
Mount Options = "rw,relatime"
Current Host  = lcars-srv-1
```

```
Resource Name = db2_regress1_0-mnt_logpath
State         = Online
Managed      = true
Resource Type = File System
Device       = "/dev/disk/by-
uuid/275f671a-d362-40d1-b4ab-ce16c34599db"
Mount Point  = "/logpath"
File System Type = ext3
Mount Options = "acl,user_xattr,noauto"
Current Host  = lcars-srv-1
```

File Systems

Qdevice Quorum

```
Fencing Information:
Configured
Quorum Information:
Qdevice
```

Qdevice information

```
-----
Model:      Net
Node ID:    1
Configured node list:
0          Node ID = 1
1          Node ID = 2
Membership node list: 1, 2
```

Qdevice-net information

```
-----
Cluster name:db2domain
QNetd host:  lcars-disk:5403
Algorithm:   LMS
Tie-breaker: Node with lowest node ID
State:      Connected
```

## Cluster Topology in this example

- A MF cluster with 2 nodes
- Each node has 1 Ethernet adapter
- Partitions active on lcars-srv-1
- 2 shared FS: one for DB, one for log
- Quorum uses Qdevice with fencing configured

# Architecture: Resource Model: *Network resource & constraints*



```
primitive db2_xela-1_eth0 db2ethmon \
  params interface=eth0 hostname=xela-1 repeat_count=4 repeat_interval=4 \
  op monitor timeout=30s interval=4 \
  op start timeout=60s interval=0s \
  op stop interval=0s timeout=20s

location no-probe-db2_xela-1_eth0-on-xela-2 db2_xela-1_eth0 resource-discovery=never -inf: xela-2
location no-probe-db2_xela-2_eth0-on-xela-1 db2_xela-2_eth0 resource-discovery=never -inf: xela-1

location prefer-db2_xela-1_eth0-on-xela-1 db2_xela-1_eth0 100: xela-1
location prefer-db2_xela-2_eth0-on-xela-2 db2_xela-2_eth0 100: xela-2
```

*db2\_xela-1\_eth0 resource is preferred to start on xela-1.  
db2\_xela-2\_eth0 resource is preferred to start on xela-2.*

- Name of the constraint
- "prefer" informs Pacemaker the preferred host of this resource (xela-1 in this case)

- The score used for comparison if this resource can be started on more than one host.

*db2\_xela-1\_eth0 resources can never start on host - xela-2 because the discovery is disabled. Vice Versa for db2\_xela-2\_eth0.*

- Name of constraint
- "no-probe" prevents Pacemaker to run the one-time monitor operation when a resource is first started.

- resource-discovery=never prevents Pacemaker to explore if this resource should be started on the host specified.
- "-inf" refers to a score of "-infinity"



# Architecture: Resource Model: *Partition resource & constraints*

Resource ID

Resource Agent

```
primitive db2_apinst1_0 db2partition \  
  params instance=apinst1 dbpartitionnum=0 \  
  op monitor timeout=120s interval=10s on-fail=restart \  
  op start interval=0s timeout=900s \  
  op stop interval=0s timeout=900s \  
  meta migration-threshold=0 target-role=Started
```

```
location lo-db2_apinst1_0-eth0-xela-1 db2_apinst1_0 \  
  rule -inf: db2ethmon-eth0 eq 0
```

```
location lo-db2_apinst1_0-eth0-xela-2 db2_apinst1_0 \  
  rule -inf: db2ethmon-eth0 eq 0
```

Ensures interface resource, eth0 on xela-1, is up (db2ethmon resource agent MONITOR action returns 0) on the xela-1 where the partition resource (db2\_apinst1\_0) is running on.

Ensures interface resource, eth0 on xela-2, is up (db2ethmon resource agent MONITOR action returns 0) on the xela-2 where the partition resource (db2\_apinst1\_0) is running on.

# Architecture: Resource Model: *Mount resource & constraints*

Resource ID

Resource Agent

```
primitive {db2_apinst1_0-mnt_db2haap} {db2fs} \  
  params device="/dev/sda" directory="/db2haap" fstype=ext3 \  
  op monitor interval=10s \  
  op_params OCF_CHECK_LEVEL=20 \  
  op start interval=0s timeout=60s \  
  op stop interval=0s timeout=60s  
  
colocation co-db2_apinst1_0-with-db2_apinst1_0-mnt_db2haap inf: db2_apinst1_0:Started  
db2_apinst1_0-mnt_db2haap:Started  
  
order or-db2_apinst1_0-mnt_db2haap-then-db2_apinst1_0 Mandatory: db2_apinst1_0-mnt_db2haap  
db2_apinst1_0
```

Colocation rule of partition resource with mount resource. This ensures both resources are started.

Order rule ensures the mount resource (specified first) starts ahead of the partition resource (specified second)

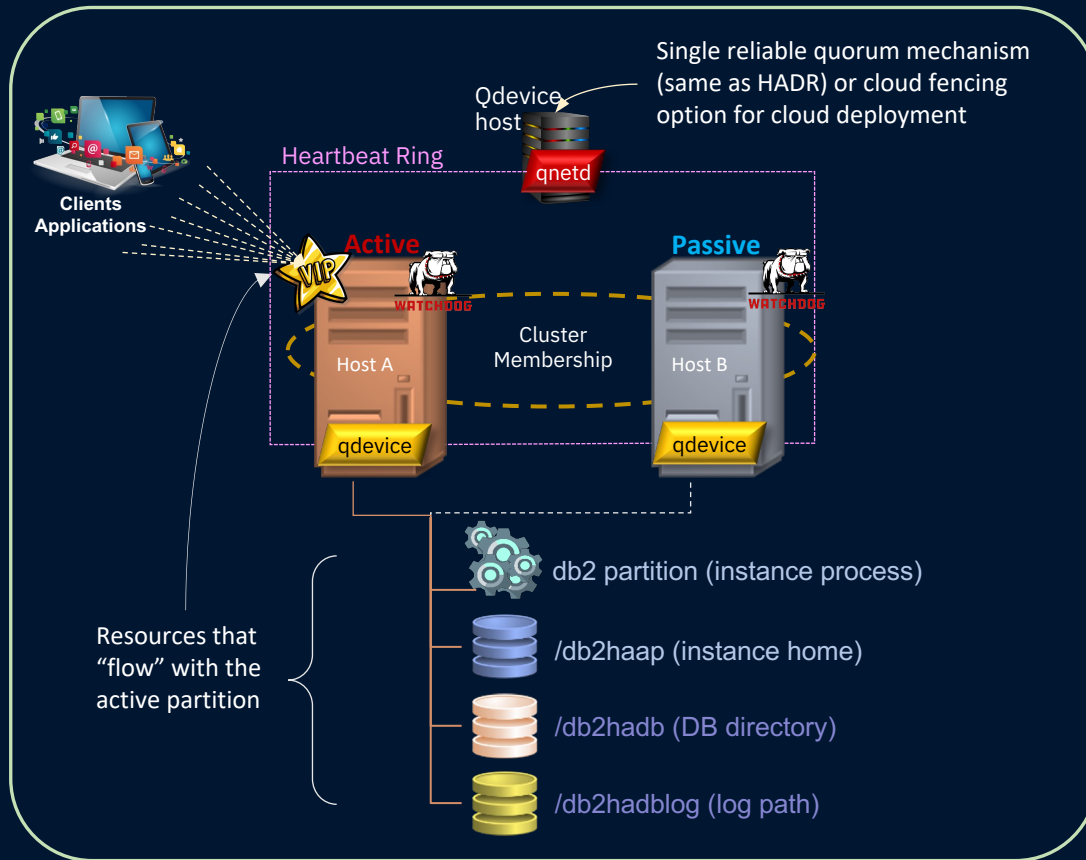
# Architecture: Resource Model: *Virtual IP resource and constraints*

```
Resource ID → Resource Agent →  
primitive {db2_apinst1_0-VIP} IPaddr2 \  
  params ip=10.31.37.222 cidr_netmask=19 \  
  op monitor interval=30s \  
  op start interval=0s timeout=20s \  
  op stop interval=0s timeout=20s  
  
colocation co-db2_apinst1_0-VIP-with-db2_apinst1_0 inf: db2_apinst1_0-VIP:Started  
db2_apinst1_0:Started  
  
order or-db2_apinst1_0-then-db2_apinst1_0-VIP Mandatory: db2_apinst1_0:start  
db2_apinst1_0-VIP:start
```

Colocation rule of VIP resource with partition resource. This ensures both resources are started.

Order rule ensures the partition resource (specified first) starts ahead of the VIP resource (specified second)

# Architecture: Resource Model: *Fencing with SBD + Qdevice*



The evicted node will be fenced via SBD service where the local watchdog reboots the node and won't be allowed to rejoin cluster until it can gain quorum.

## Why is fencing mandatory with MF ?

- Shared storage
- Need to handle potential strayed process(es) in split brain scenario and host failure

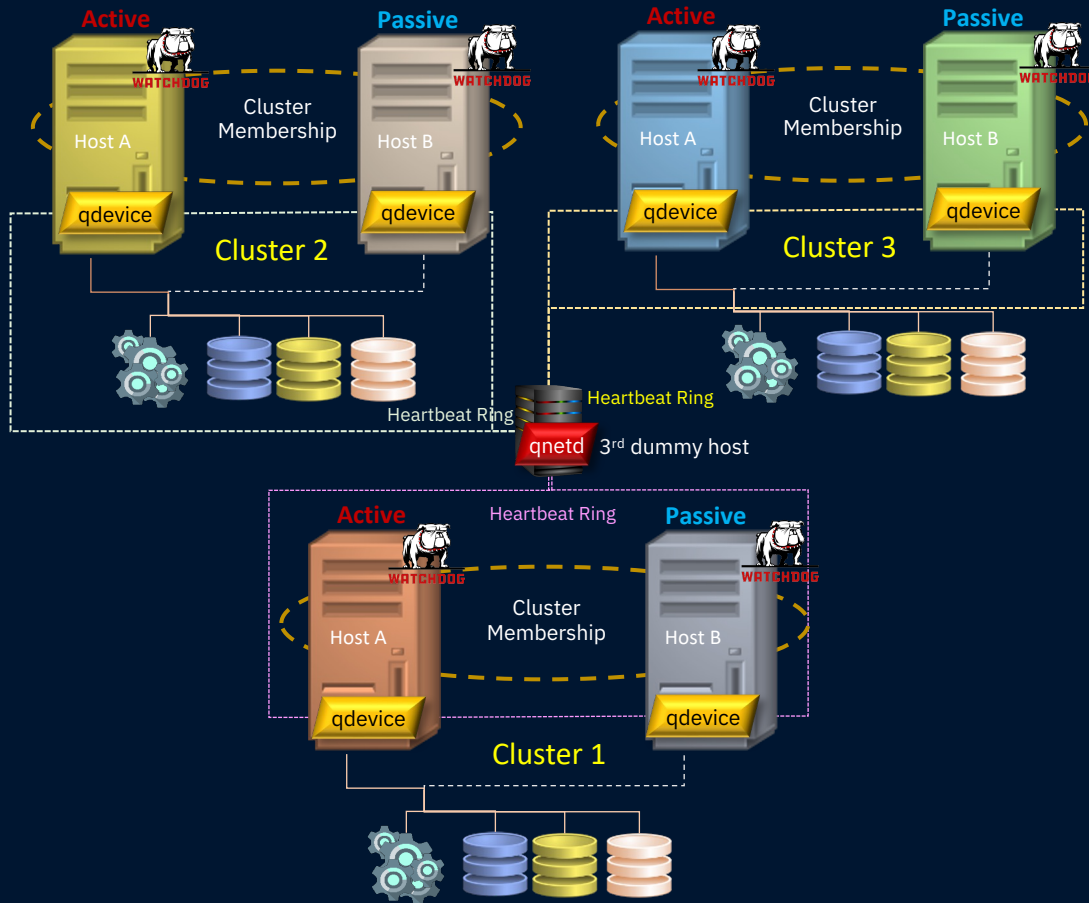
## How does it work ?

- Diskless SBD (STONITH Block Device) provides a node fencing mechanism via reboots when quorum is lost without the need of shared disk
- SBD and watchdog (Hardware or Software) must be configured for fencing to work.
- Software watchdog setup is handled by Db2 install if hardware watchdog is not configured.

## When is fencing triggered ?

1. Quorum Loss. Communication loss between cluster nodes. Qdevice votes for the node with the lowest node ID.
2. An attempt to stop a mount resource fails

# Architecture: Resource Model: *Quorum with Qdevice (3<sup>rd</sup> host)*



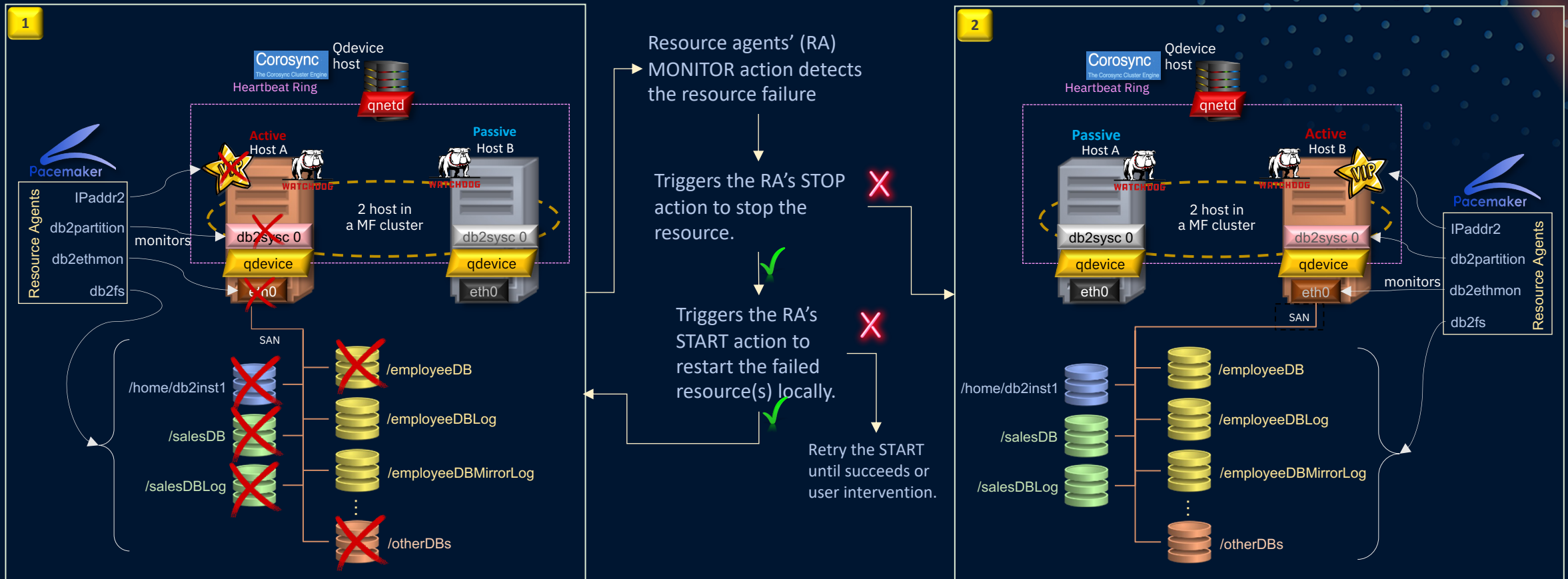
## Setup

- **qdevice** - separate daemon from Pacemaker running on each cluster nodes
- **qnetd** - standalone daemon running on a 3<sup>rd</sup> host (not in cluster)
- TCP/IP connectivity among the 3 processes

## 3<sup>rd</sup> host detail

- Flexible in platform, architecture
- TCP/IP accessible from all hosts
- Possible to share with other Pacemaker clusters
  - e.g. use a RHEL host on Z for clusters nodes on POWER RHEL, x86 SLES, and Z with SLES.
- Small memory, disk footprint
  - Only need to install corosync-qnetd RPM
  - No need to install Db2 or Pacemaker
  - Not part of the Pacemaker cluster

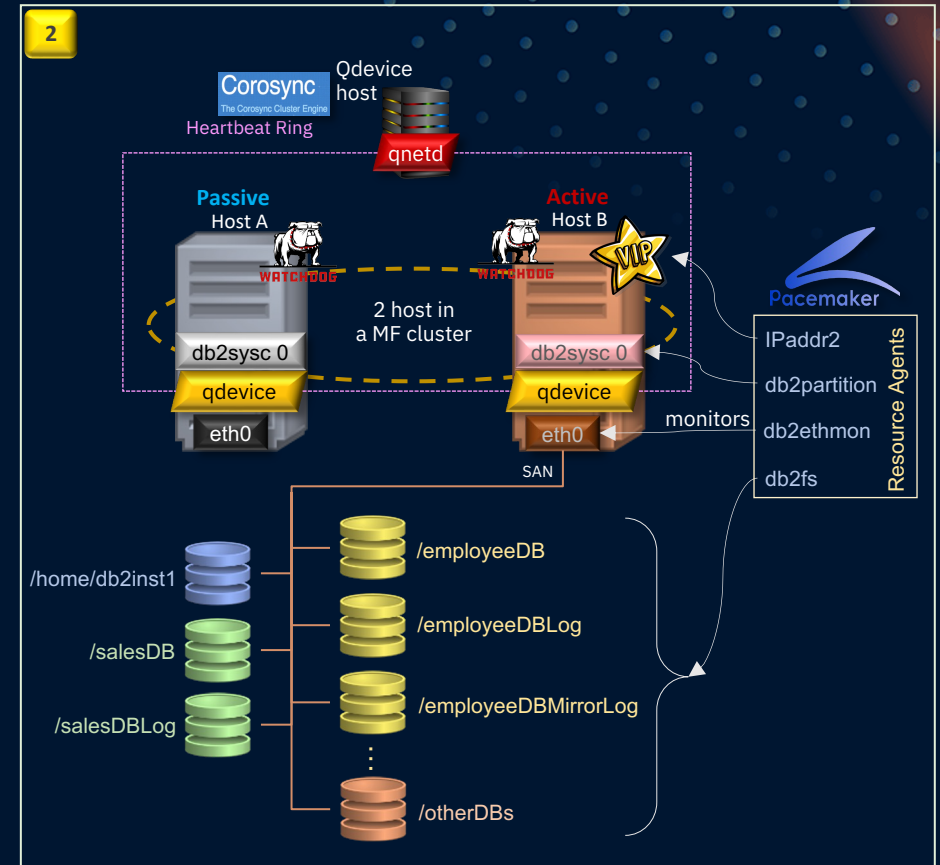
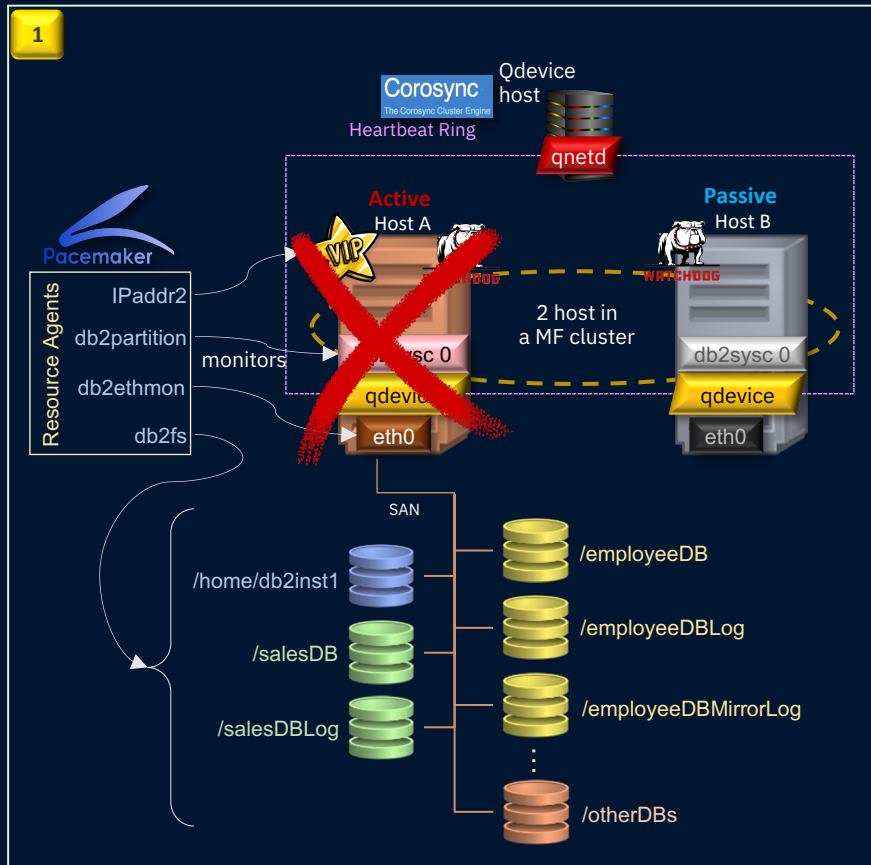
# Failure Behaviour: *Resource Failure*



## Results

- Resource failure leads to local restart of the resource
- Fencing only occurs if the failed resources failed to be stopped by Pacemaker.

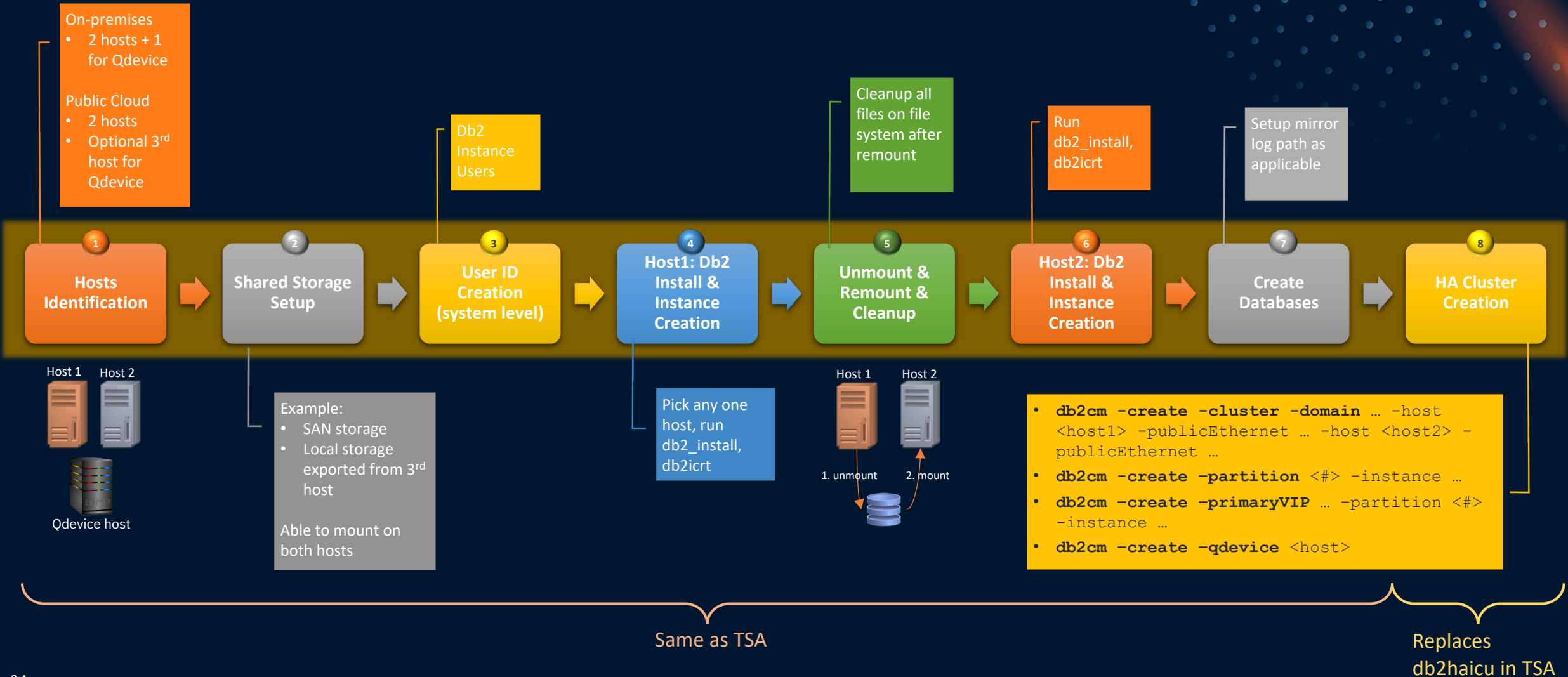
# Failure Behaviour: *Host Failure*



## Results

- Corosync detects loss of quorum on HostA, notify Pacemaker to restart all resources on the other hosts

# Up & Running End-to-End Overview





# db2cm options

## Up & Running

- `db2cm -create -cluster -domain <name> -publicEthernet <NIC> -host <host1> -publicEthernet <NIC> -host <host2>`
- `db2cm -delete -cluster`
- `db2cm -create -partition <#> -instance <instname>`
- `db2cm -delete -partition <#> -instance <instname>`
- `db2cm -create -primaryVIP <IPv4 addr> -partition <#> -instance <instname>`
- `db2cm -delete -primaryVIP -partition <#> -instance <instname>`
- `db2cm -create -qdevice <host>`
- `db2cm -delete -qdevice`
- `db2cm -export <filename>`
- `db2cm -import <filename>`

## Maintenance

- `db2cm -create -mount <mount point> -partition <#> -instance <instname>`
- `db2cm -delete -mount <mount point> -partition <#> -instance <instname>`
- `db2cm -add -dbMount <DB Name> -partition <#> -instance <instname>`
- `db2cm -remove -dbMount <DB Name> -partition <#> -instance <instname>`
- `db2cm -disable <-all | -partition <#> -instance <instname>>`
- `db2cm -enable <-all | -partition <#> -instance <instname>>`
- `db2cm -set -option mountMonitoring <mountPoint> -value <yes|no> -partition <#> -instance <instname>`
- `db2cm -move -partition <#> -instance <instname> -host <target hostname>`

## Monitor

- `db2cm -list`

# Sequence of commands to setup a Mutual Failover cluster

## 1. Create Cluster

```
db2cm  
-create -cluster  
-domain <name>  
-publicEthernet <e>  
-host <host>  
-publicEthernet <e>  
-host <host>
```

## 2. Create partition resource

```
db2cm  
-create  
-partition <#>  
-instance <name>
```

## 3. Create Qdevice quorum

```
db2cm  
-create  
-qdevice <host>
```

## 4. Create DB mount resources

```
db2cm  
-add  
-dbMount <MP>  
-partition <#>  
-instance <name>
```

Optional  
only

### Create VIP resource

```
• db2cm -create -primaryVIP 170.120.1.1  
-netmask 21 -partition 0 -instance  
db2inst1
```

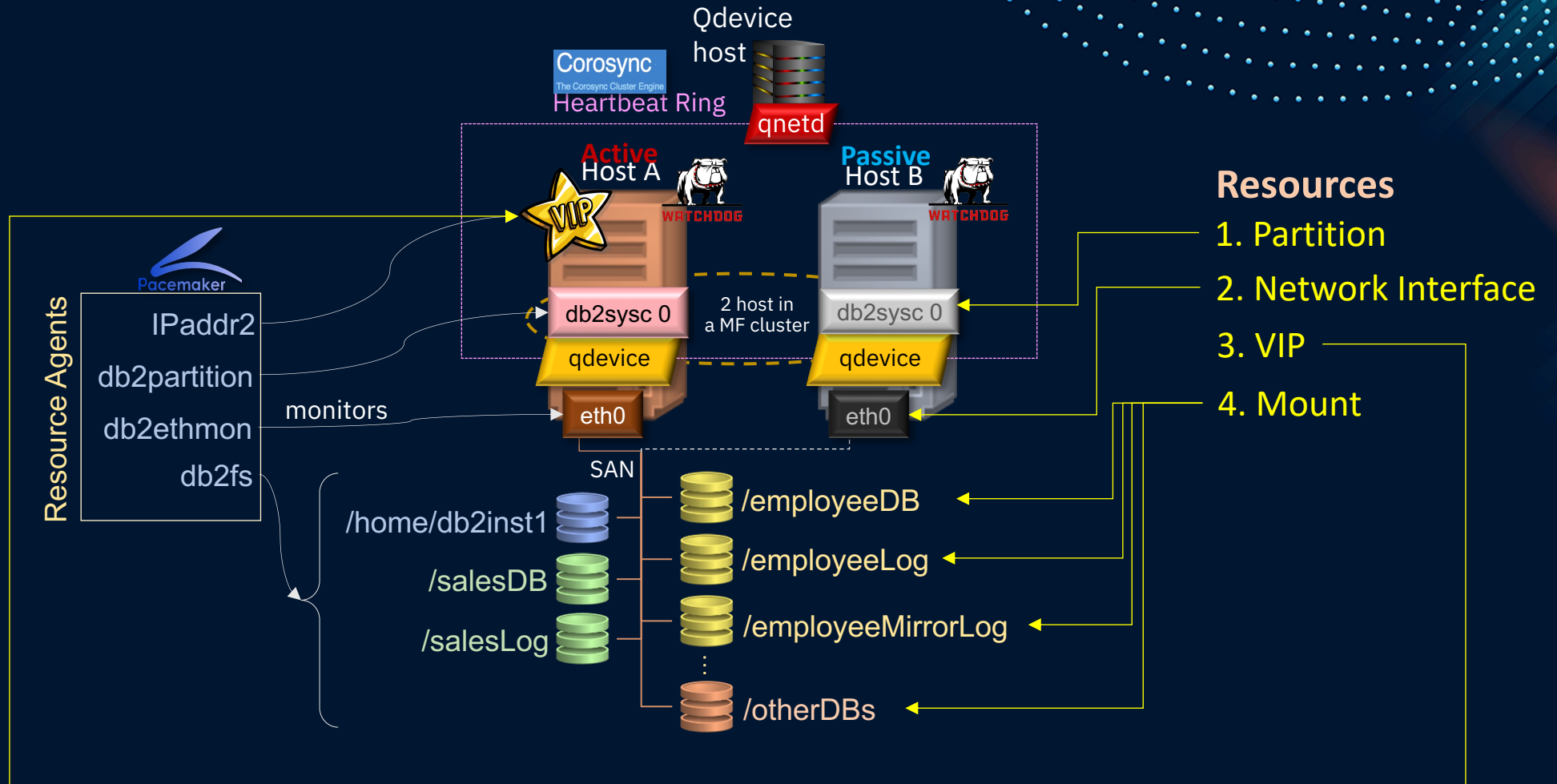
### Create extra DBs' resources

```
• db2cm -add -dbMount employee_db2  
-partition 0 -instance db2inst1
```

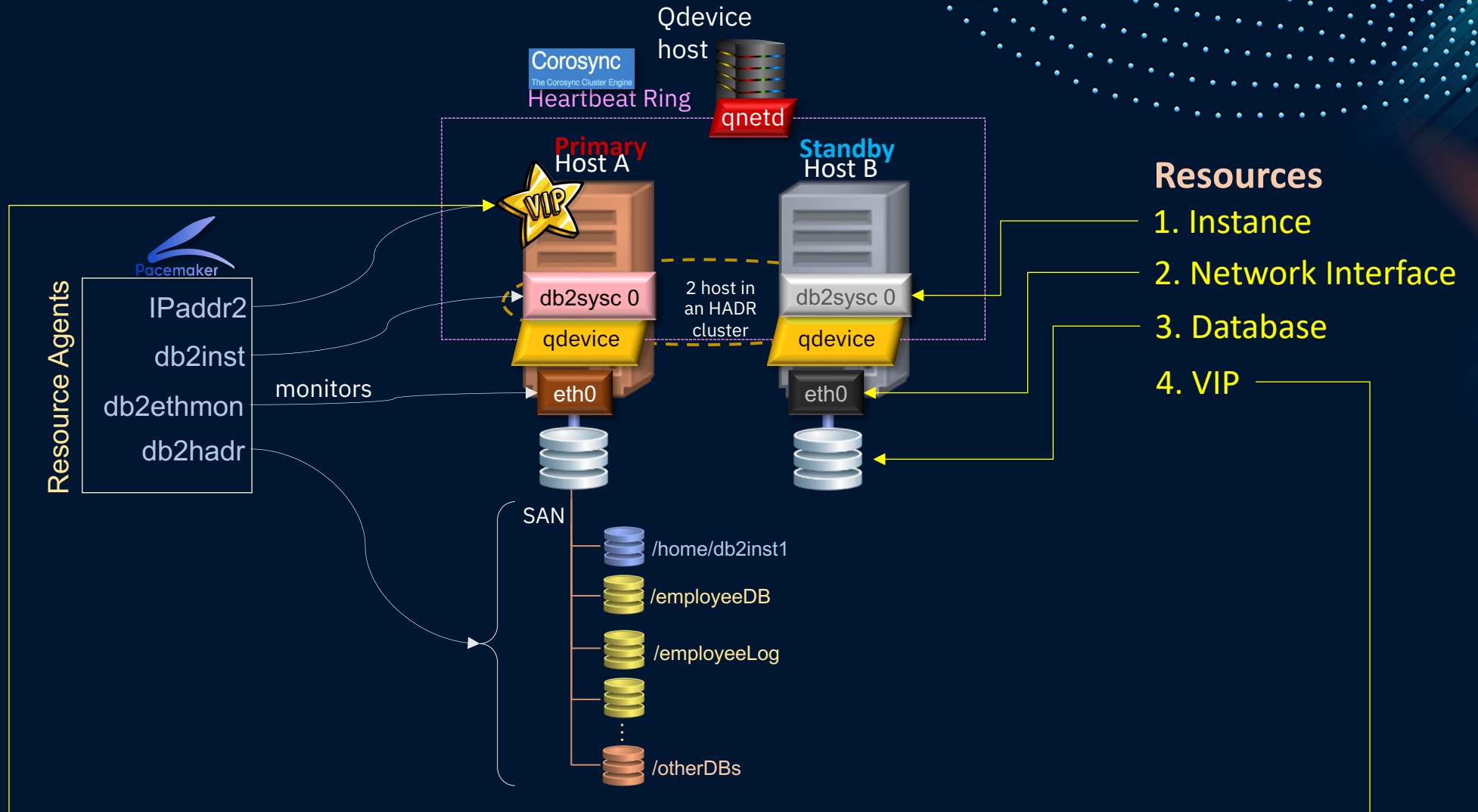
### Create other mount resource

```
• db2cm -create -mount disk2 -partition 0  
-instance db2inst1
```

# Mutual Failover

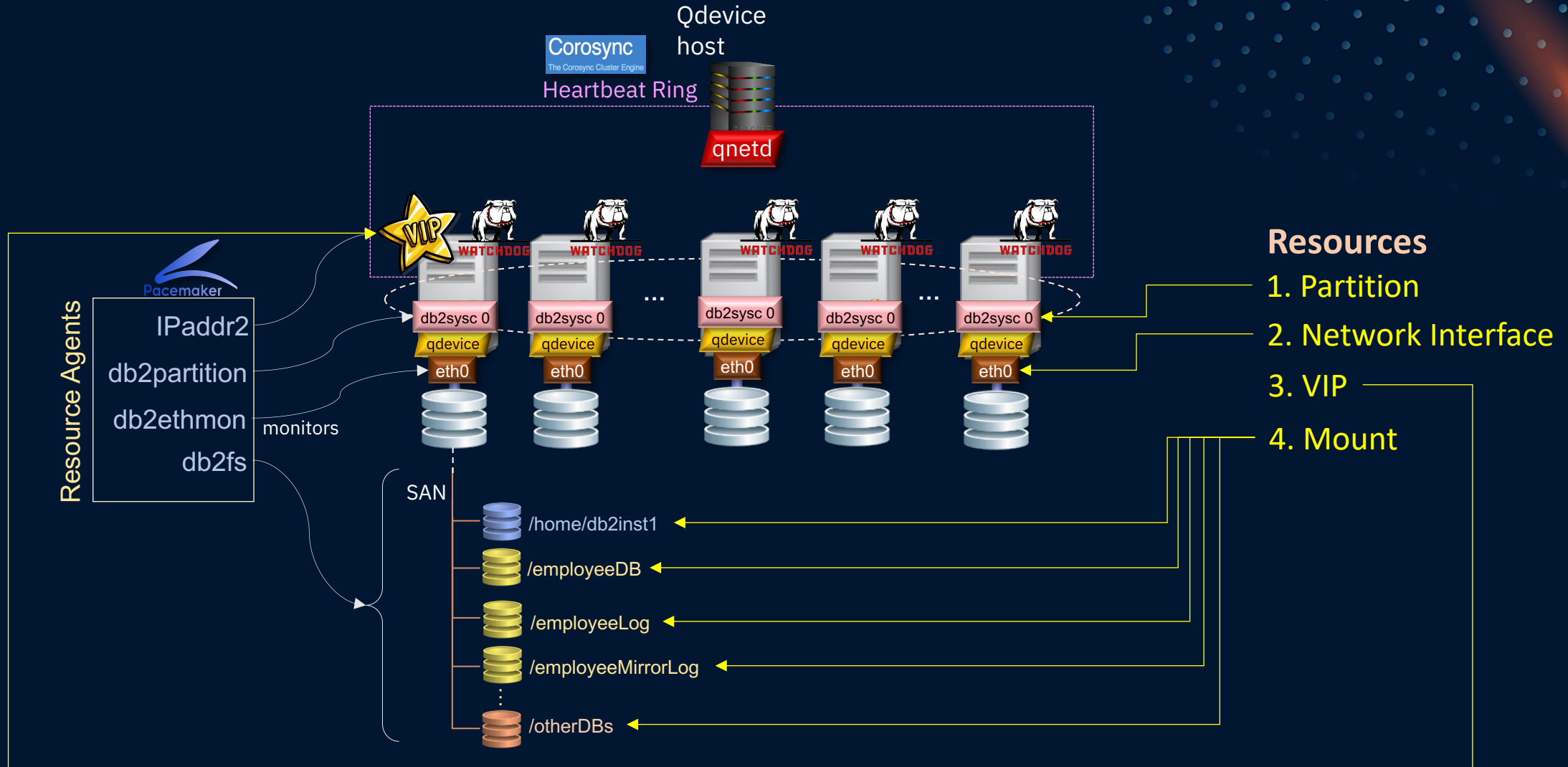


# HADR

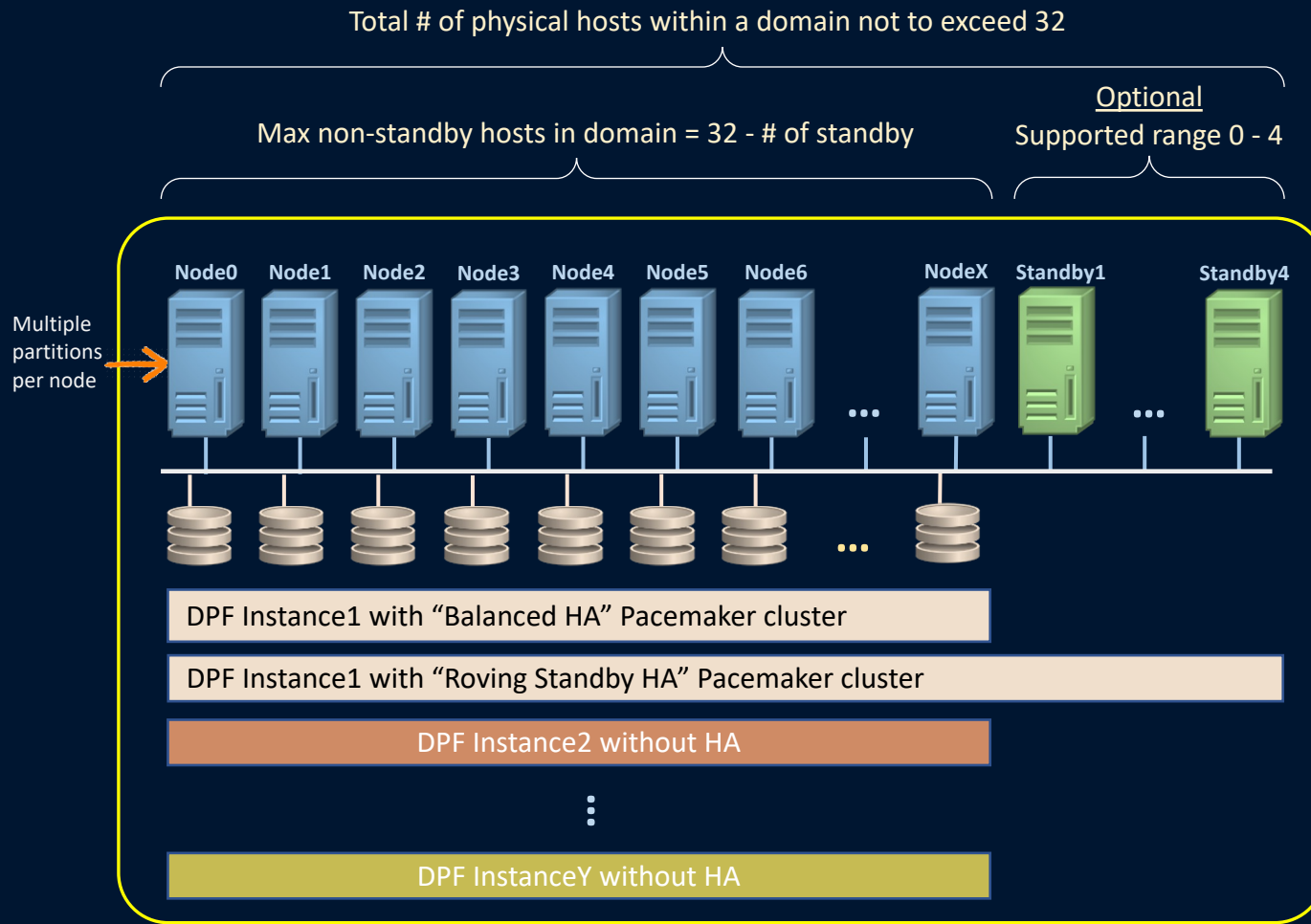


**NEW**

# Database Partitioning Feature (DPF)



# Sneak Peak at DPF HA topology with Pacemaker



## Single Pacemaker domain with one of the following failover policies:

1. Balanced HA – without standby host
2. Roving Standby HA - 1 to 4 standby host(s)
  - Provide up to 4 concurrent host failure

## Multiple instances is supported but ...

- Only one instance can have HA enabled.
- All instances can span across all hosts, but only the HA enabled instance can use the standbys

## Max number partitions supported

Using rule of thumb of 8 partitions per physical hosts:

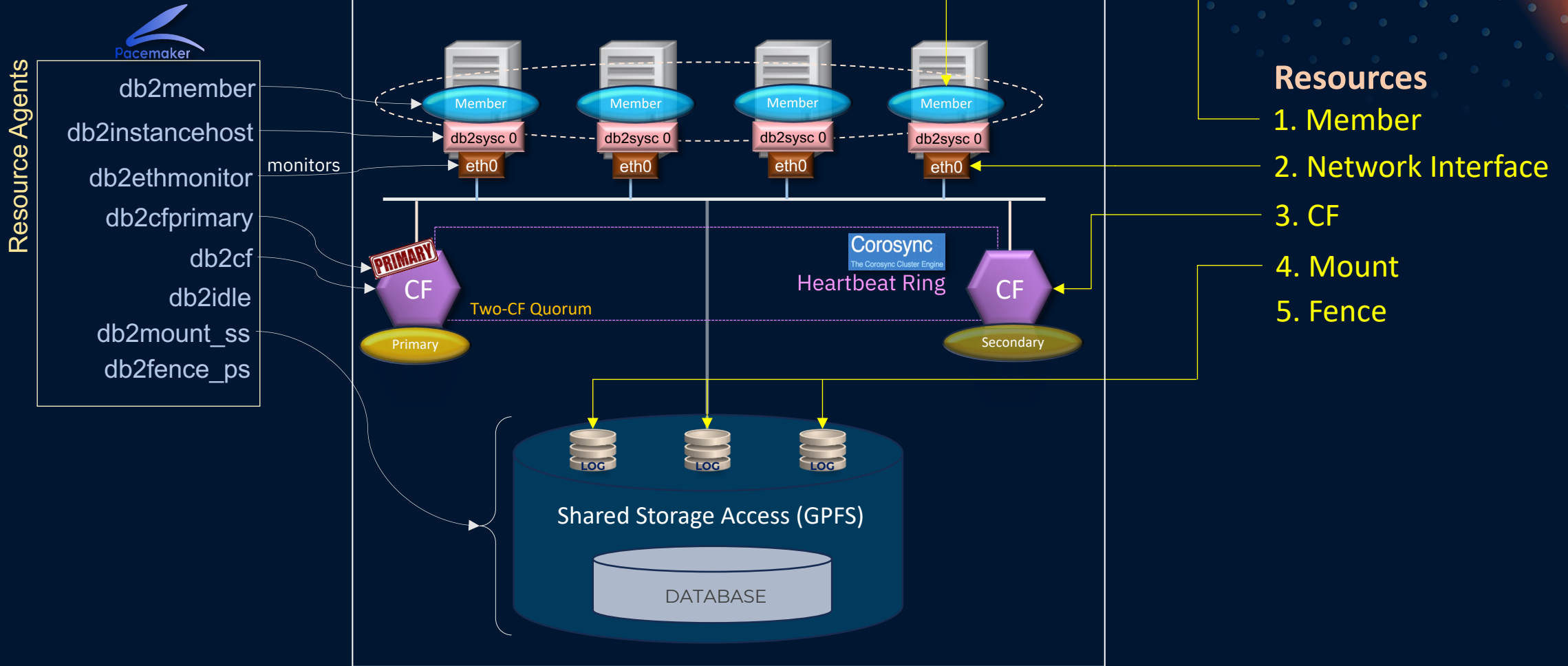
- Balanced HA: 8 per host \* 32 hosts = **256**
- Roving Standby HA: 8 per host \* (32 - 4) hosts = **224**

## Note:

- Higher number of partitions can explore deploying more partitions per host than 8 with proper H/W



# pureScale





# V12.1 Highlights with Pacemaker

DPF HA

- THIRD HA configuration with Pacemaker!

pureScale HA

- FOURTH HA configuration with Pacemaker!

pureScale on AWS

- Tech preview for pureScale on AWS with EFA

Pacemaker Refresh

- Upgrade to latest release – 2.1.8

Support newer OS level

- Validated on RHEL 9.4

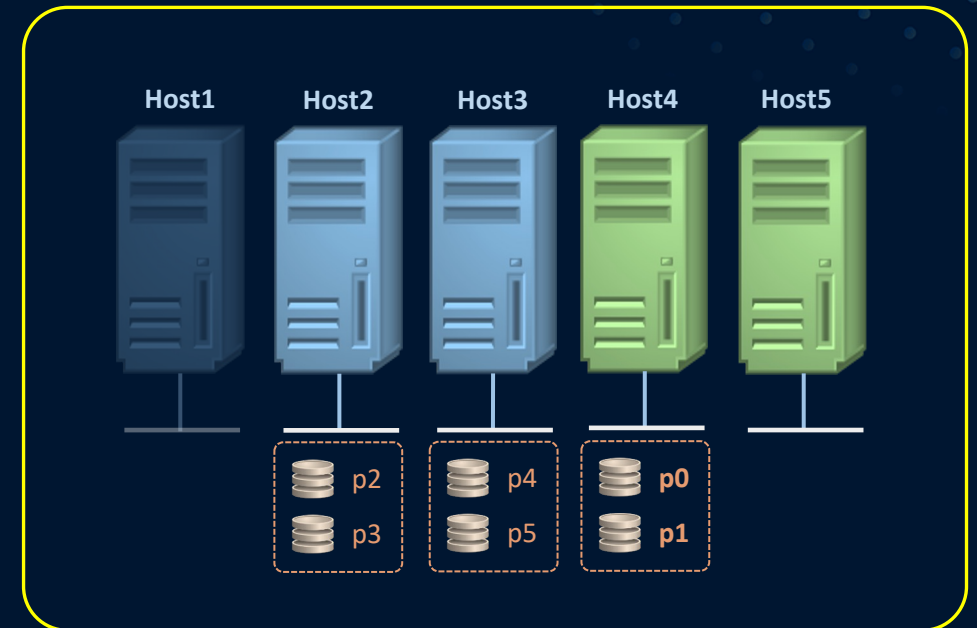
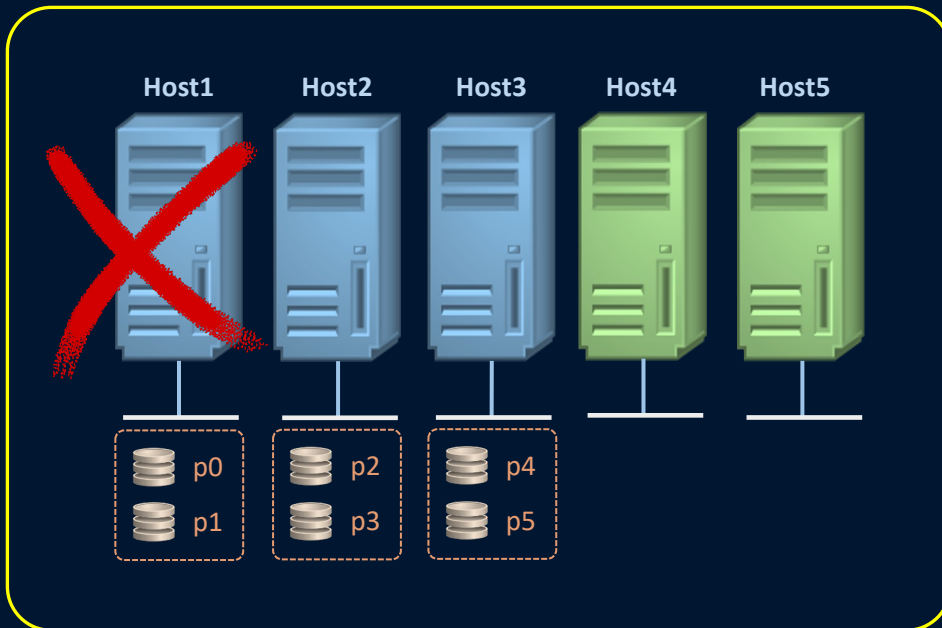


# DEMO

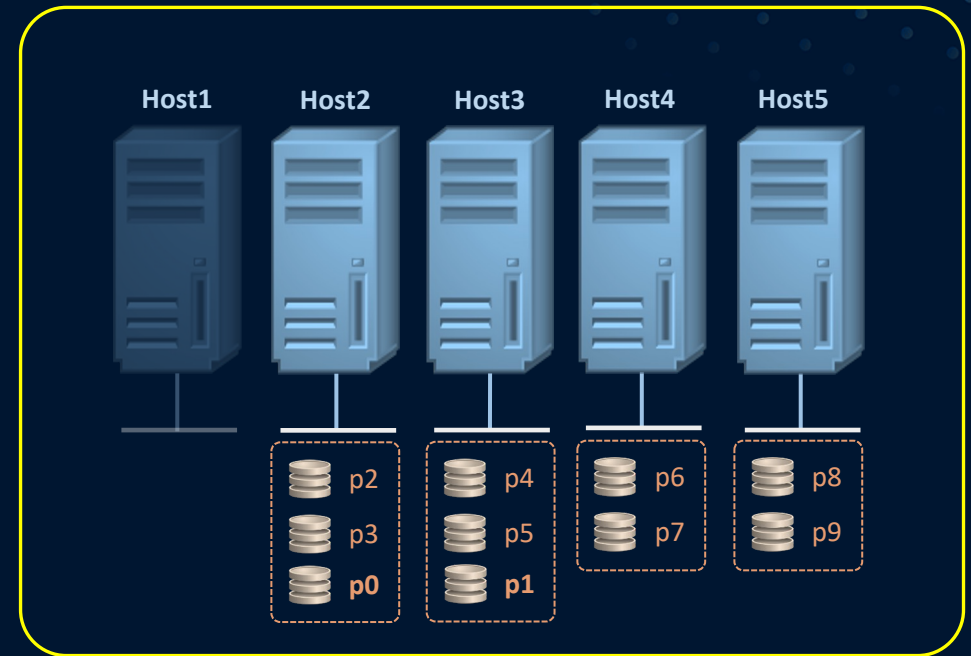
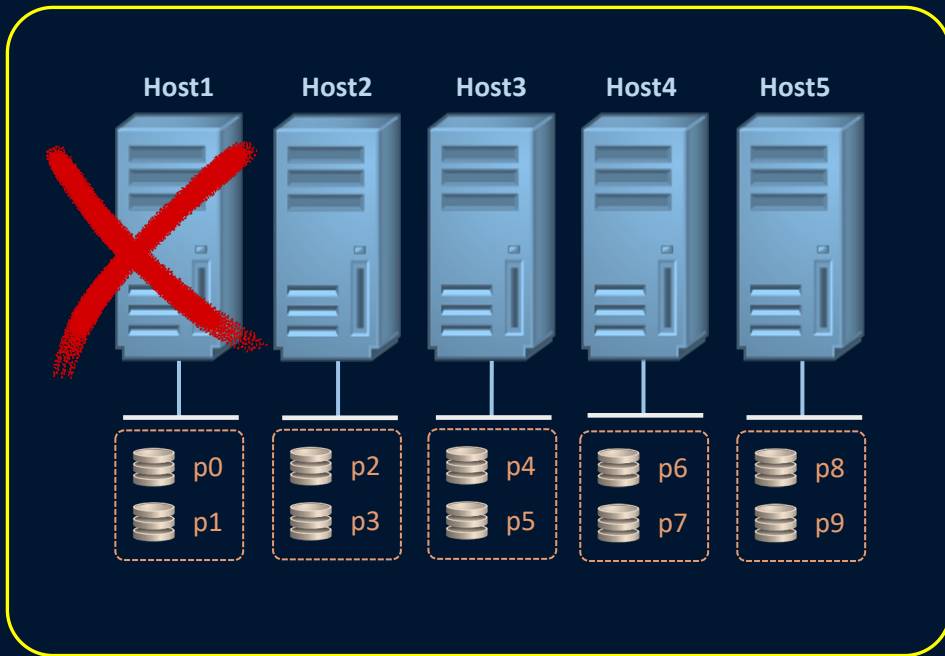
---

Showcasing DPF automated  
failover with Pacemaker

# DPF Roving Standby – Failover Behaviour



# DPF Balanced HA – Failover Behaviour





THANK  
YOU