TRIDEX

September 2023

The Transition of Db2 Automated HA from TSA to Pacemaker - *Into-Act IV*

Dr. Toby Haynes PhD

IBM Canada Ltd.

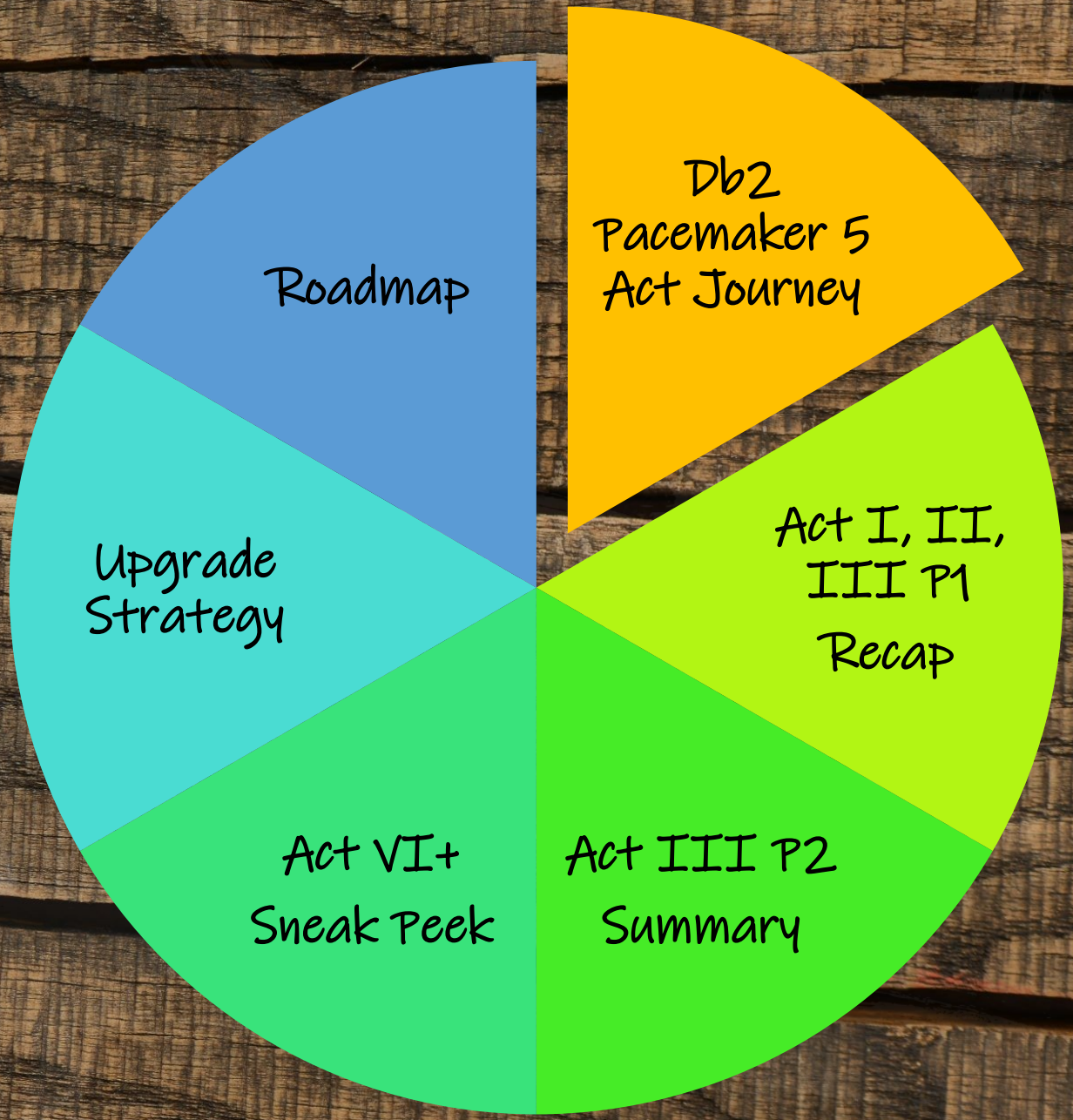
Platform: UNIX, Linux

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

Today's AGENDA



UPDATED

Db2 Pacemaker 5 Act Journey

Prologue

Key Plot

Inaugural release of cloud-ready HADR solution with Pacemaker on Linux as technical preview

Act I

Key Plot - GA

Support Production Deployment on any cloud and on-premise x86 and Z Linux environments

Climax #1

Act II

Key Plot

Integrated Bundling and Automatic Installation of Pacemaker

Act III Part 1

Key Plot

Complete HADR on all Linux architectures

Develop multiple subplots …

Act III Part 2

Key Plot

Cloud-ready 2 node Mutual Failover with shared disk

More enhancements on Cloud

Act IV

Key Plot

Declare Independence on Linux from TSA

Other HA configs

Climax #2

Act V

Key Plot

Major contribution to open-source community with Pacemaker support on AIX

Climax #3

Epilogue

Key Plot

Closure on other on-going plot lines (aha ideas)

Subplot 2 – Mutual Failover

Subplot 3 – DPF

Subplot 4 - pureScale

Subplot 5 – container support

V11.5.4.0 | V11.5.5.0 | V11.5.6.0 | V11.5.7.0 | V11.5.8.0 | Next | Next | Next

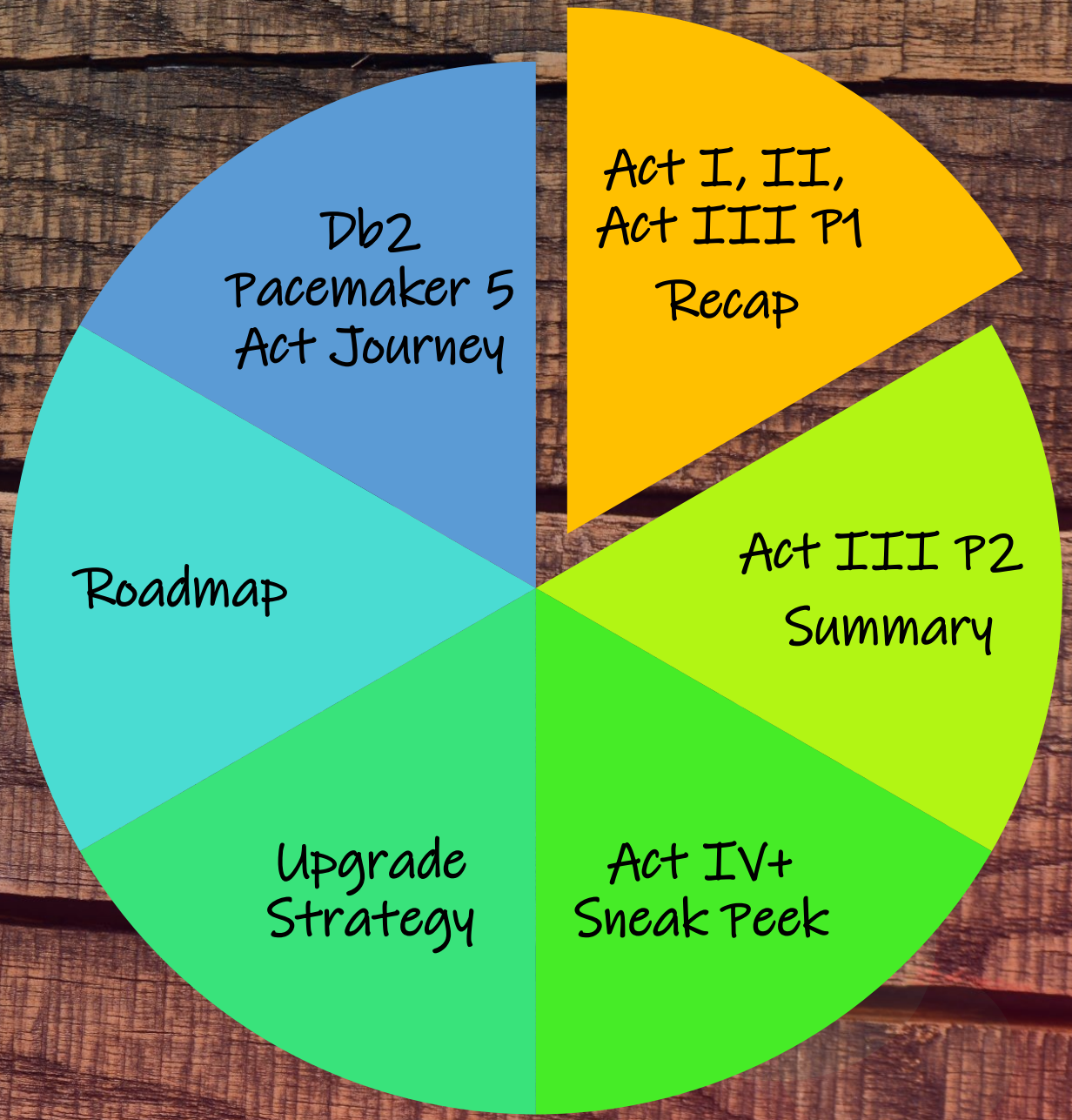
Note: Planned acts may be delayed, new acts may be added without notice

Our vision with 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 next major release

Today`s AGENDA



HADR

Capabilities Comparison with TSA through completion of Acts I, II, III

Comparable

- Platforms & H/W Architectures support
- Bundled solution
- Integrated install
- Fast redeployment
- Advanced HADR DB hang detection

Superior

- Cloud-ready
- Faster Recovery Performance
- Db2-aware network resiliency
- Higher quality with more comprehensive test scenario
- Faster support & PD
- Improved Documentation
- Direct impact on Pacemaker

Supported Platforms, Environment and HA Configurations

Comparable

- Platforms & H/W Architectures support
- Bundled solution
- Integrated install
- Fast redeployment
- Advanced HADR DB hang detection

Take Note

- *No plan to support Pacemaker as Integrated solution with older version of RHEL (7.x) and SLES (12 SPx)*
- *No plan to support Pacemaker as integrated solution in earlier Db2 releases than 11.5.5.0*

Categories	Descriptions			TSA	Pacemaker
Architecture / Platforms / OS Version	Intel	RHEL	8.x	Yes	Yes since 11.5.5.0+
	Intel	SLES	15 SPx	Yes	Yes since 11.5.5.0+
	Linux on IBM Z	RHEL	8.x	Yes	Yes since 11.5.5.0+
	Linux on IBM Z	SLES	15 SPx	Yes	Yes since 11.5.5.0+
	POWER	RHEL	8.x	Yes	Yes since 11.5.7.0+
	POWER	SLES	15 SPx	Yes	Yes since 11.5.7.0+
	POWER	AIX	7.2, 7.3	Yes	Future
Environment	On-premise DC			Yes	Yes since 11.5.5.0+
	Non-containerized Private Cloud			No	Yes since 11.5.5.0+
	Non-containerized Public Cloud			No	Yes since 11.5.5.0+
	Container			No	Future
Supported HA configurations	HADR with Multiple Standby			Yes	Yes since 11.5.5.0+
	2-node Mutual Failover with shared disk			Yes	Yes since 11.5.8.0+ NEW!
	DPF HA			Yes	In development
	pureScale			Yes	In development

Take Note

New from 11.5.7.0 and up: support statement has been relaxed from specific RHEL & SLES version to any newer ones within same major release (*i.e. RHEL 8.x, SLES 15 SPy*)

Up & Running Improvements

Comparable

- Platforms & H/W Architectures support
- Bundled solution
- Integrated install
- Fast redeployment
- Advanced HADR DB hang detection

Db2 11.5.5.0

- Separate download of Pacemaker from Marketing Registration Site (MRS)

Db2 11.5.6.0

- Pacemaker bundled with Db2
- Integrated install via command line - db2_install

Db2 11.5.8.0 NEW

- Integrated install via silent install & GUI
- Automated setup fencing & VIP on cloud

More on this later …

db2cm (link)

-export

Backup the cluster configuration to a file which can be used with the `-import` option to quickly redeploy the cluster on the same set of hosts.

For example:

```
sqllib/bin/db2cm -export /tmp/backup.conf
```

For more details refer to Backup cluster configuration information.

-import

Restore the cluster configuration from a previously saved configuration generated by the `-export` option.

For example:

```
sqllib/bin/db2cm -import /tmp/backup.conf
```

For more details refer to Restore from a saved Pacemaker cluster configuration.

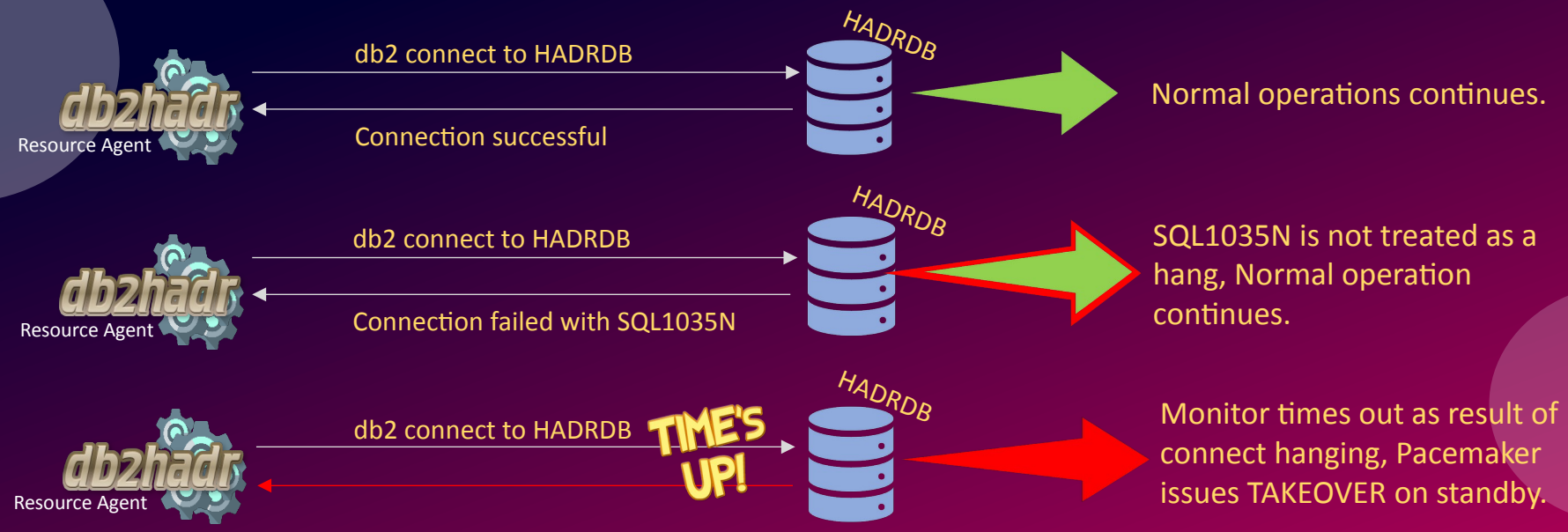
Works for both HADR and Mutual Failover (new in 11.5.8.0)

Advanced HADR DB hang detection on Linux (Pacemaker view)

Resource agent - *db2hadr* supports detecting hangs while connecting to the primary database.

Comparable

- Platforms & H/W Architectures support
- Bundled solution
- Integrated install
- Fast redeployment
- Advanced HADR DB hang detection



Enablement:

- Off by default, enabled via environment variable. Effectively immediately, no instance restart required.
- Add the following to instance user's $HOME/.profile

```
export DB2_HADR_HANG_DETECTION=CONNECT
```

To bypass specific SQLN codes:

- `export DB2_HADR_HANG_SQL_BYPASS=SQL1040N,SQL1035N,SQL1060N`
- Ignored codes will not result in the monitor returning a failed state (so no TAKEOVER issued)

Cloud-Ready with AWS & Azure VIP support

Superior

- ✓ Cloud-ready
- Faster Recovery Performance
- Db2-aware network resiliency
- higher Quality by having more comprehensive test scenario
- Faster support & PD
- Improved Documentation
- Direct impact on Pacemaker

Db2 11.5.5.0

- Cloud-Ready
- Alternate validated config on AWS

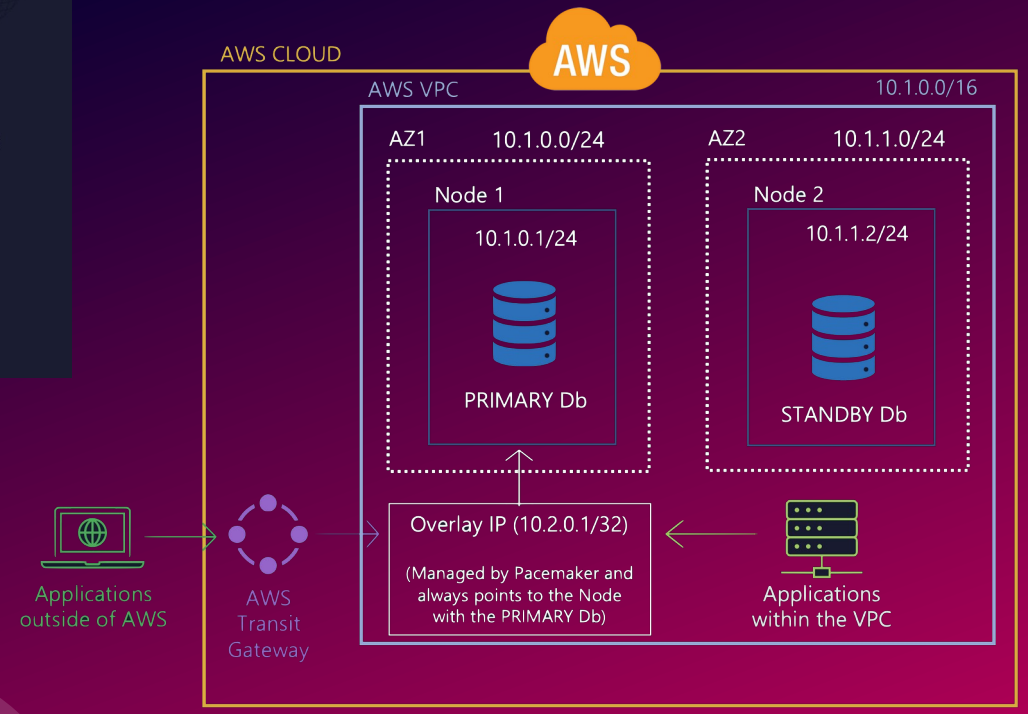
Db2 11.5.6.0

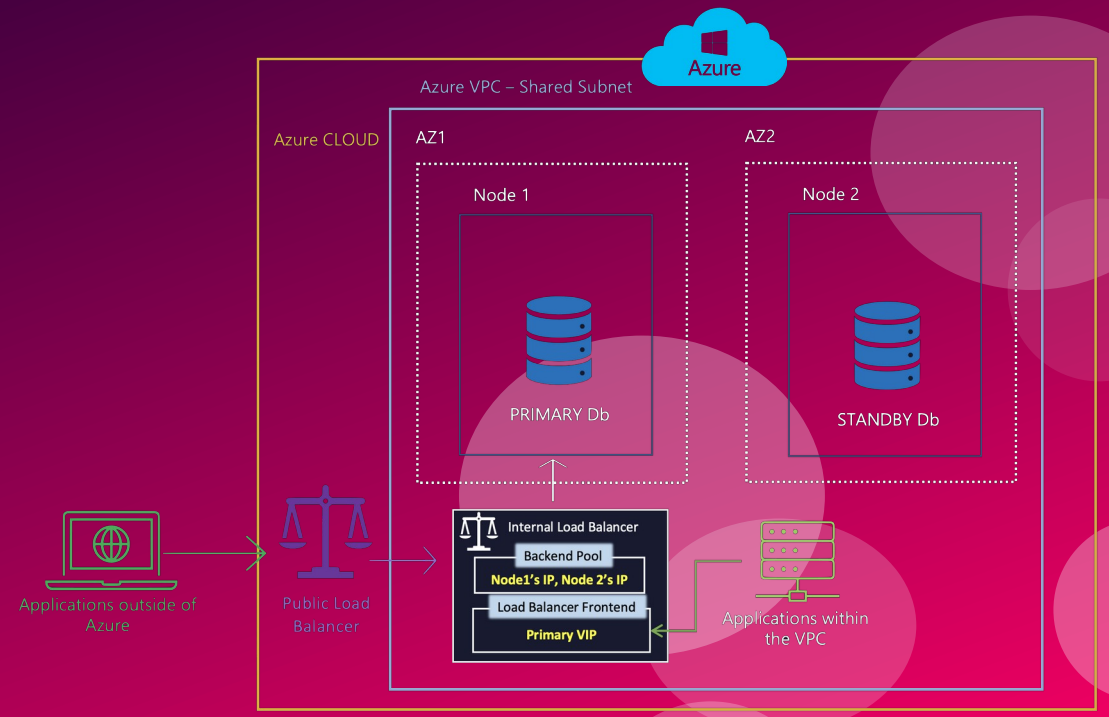
- Alternate validated config on Azure

Db2 11.5.8.0 NEW

- Automated setup on AWS & Azure*

More on this later …





Cloud-Ready with AWS & Azure Fencing Support

In lieu of 3rd host for QDevice Quorum

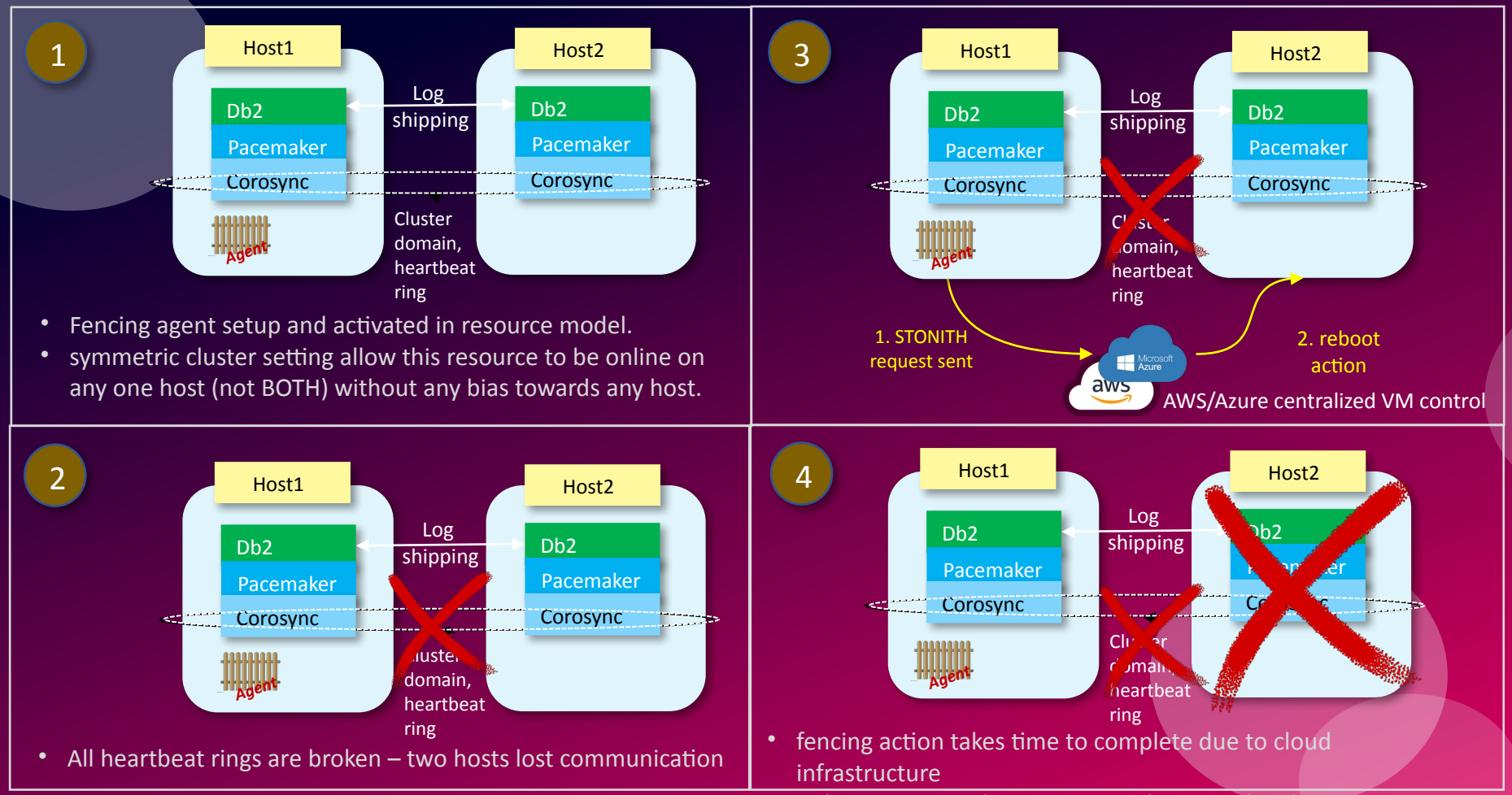
Fence agent available via the IBM hosted – Market Registration Site (MRS)

Superior

- Cloud-ready
- Faster Recovery Performance
- Db2-aware network resiliency
- higher Quality by having more comprehensive test scenario
- Faster support & PD
- Improved Documentation
- Direct impact on Pacemaker

Qdevice Vs Fencing ?

RULES



1
- Fencing agent setup and activated in resource model.
- symmetric cluster setting allow this resource to be online on any one host (not BOTH) without any bias towards any host.

2
- All heartbeat rings are broken – two hosts lost communication

4
- fencing action takes time to complete due to cloud infrastructure
- In-house testing shows up to 6x slower with Qdevice

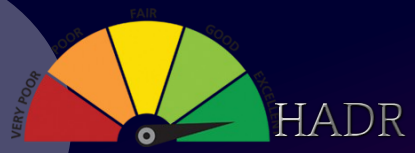
Based your decision on the need for faster recovery from primary host failure Vs on-going cost of maintaining a small 3rd VM. (can be as low as $53.00 USD/month (AWS EC2 t4g.large instance, 2 vCPU, 8G memory, 30G EBS, up to 5Gbps network)

Price is subjected to change without notice

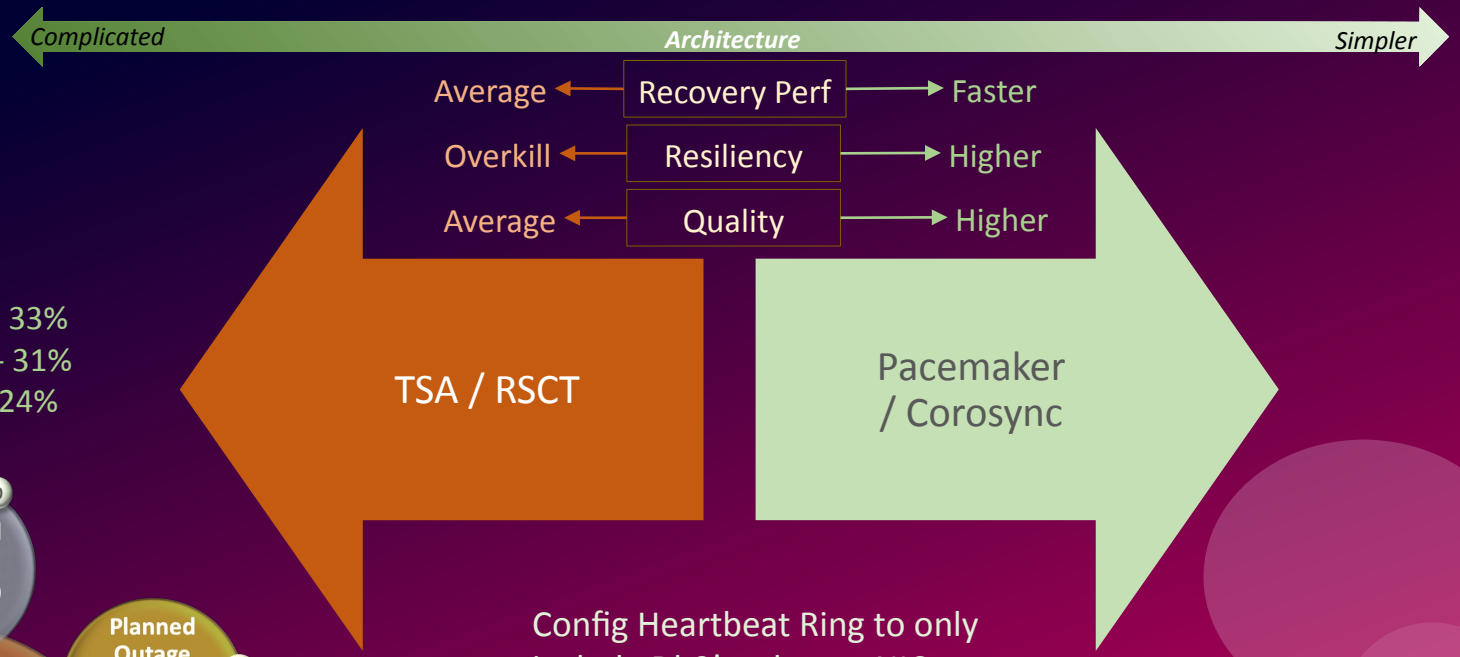
Resiliency, Recovery Performance, Quality

Superior

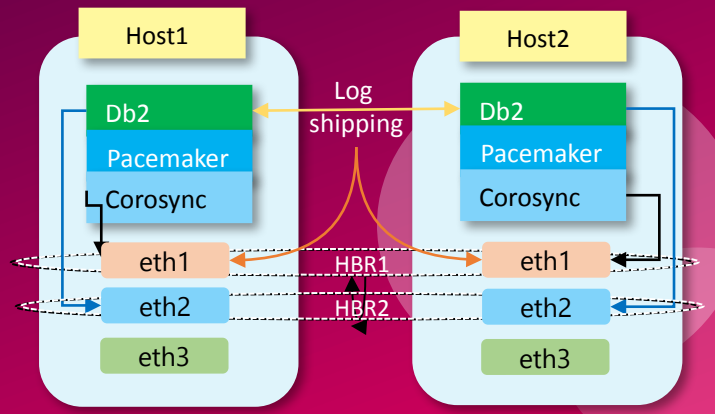
- Cloud-ready
- ✓ Faster Recovery Performance
- ✓ Db2-aware network resiliency
- ✓ higher Quality by having more comprehensive test scenario
- Faster support & PD
- Improved Documentation
- Direct impact on Pacemaker



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



Config Heartbeat Ring to only include Db2's relevant NICs



Streamlined Support, Self-diagnosis Possible!

Superior

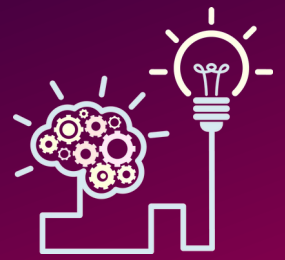
- Cloud-ready
- Faster Recovery Performance
- Improved Documentation
- higher Quality by having more comprehensive test scenario
- Db2-aware network resiliency
- Faster support & PD
- Direct impact on Pacemaker

Complicated — ***Architecture*** — *Simpler*

RecoveryRM, ConfigRM leaders
Circular logs, spool logs, etc.

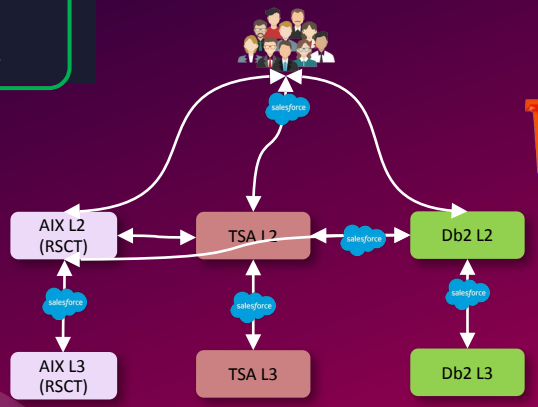
TSA / RSCT

Pacemaker / Corosync



/var/log/pacemaker/pacemaker.log
/var/log/cluster/corosync.log

Lost count!	# of logs/host	2
Binary	Log file format	Text
Impossible	Self-diagnosis	Very doable
Slower	Case support	Faster

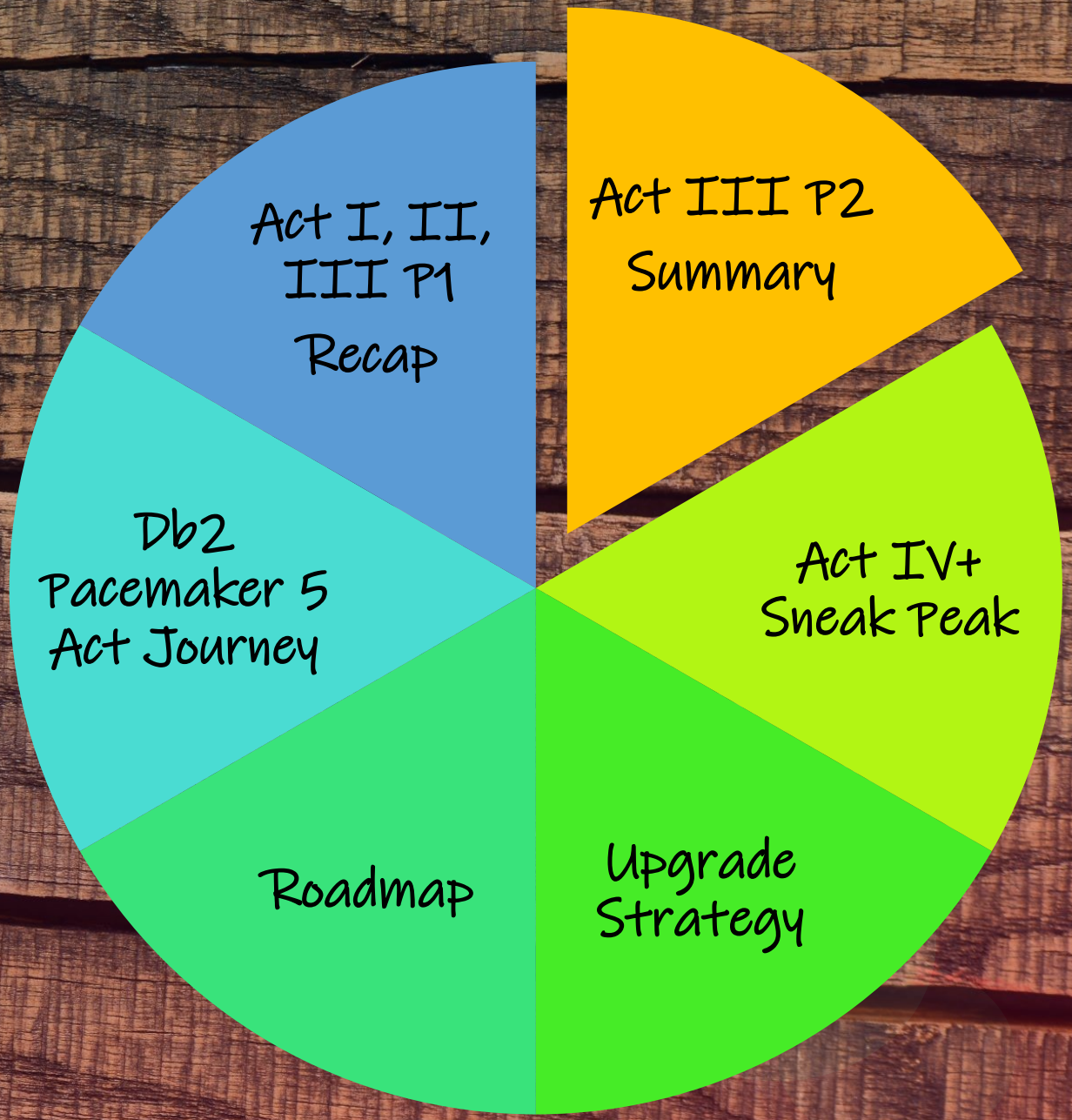


VS

IBM DB2 contributes → Pacemaker open source

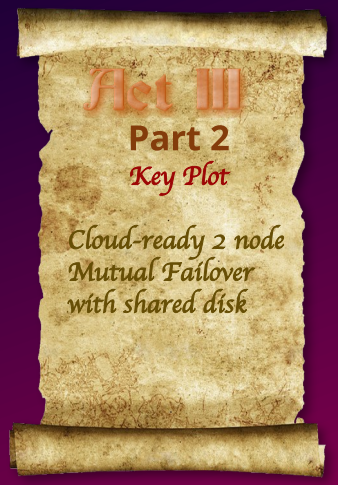
First contribution made from Db2 development to the open-source Pacemaker community

Today's AGENDA



NEW

V11.5.8.0 Highlights with Pacemaker

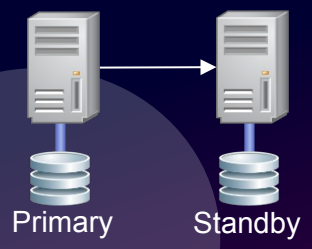


Mutual Failover HA	• *SECOND* HA configuration with Pacemaker!
Expanded integrated installation methods	• Integrated Silent & GUI Pacemaker installations
Cloud Enhancements	• New options to drastically reduce number of manual steps.
Pacemaker Refresh	• Upgrade to latest release – 2.1.2
Support newer OS level	• Validated on RHEL 8.6

A 10,000' look at Db2 Cloud-Ready *Integrated* HA Topologies with Pacemaker

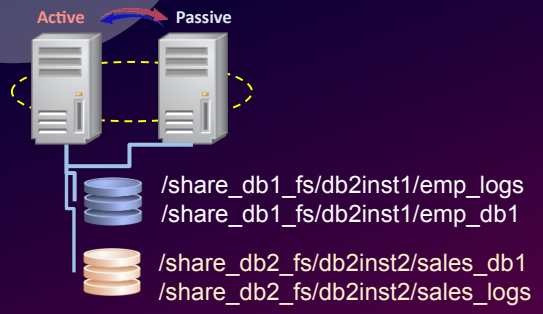
V11.5.5.0

Single DB Partition (EE) with automated HADR



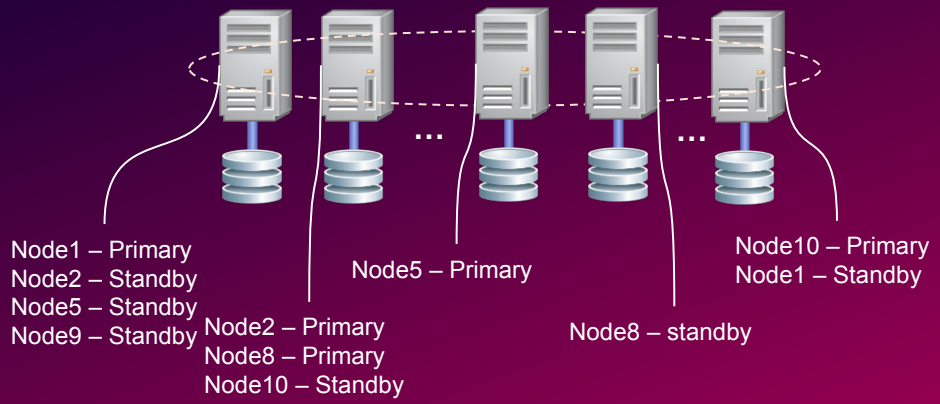
V11.5.8.0

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



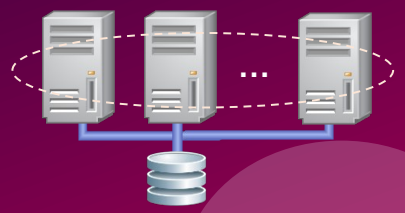
FUTURE

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



FUTURE

pureScale Online 24x7x365 with automatic failover



Highlights:

- Cloud-ready
- Unlike HADR with independent storage in each host, MF supports shared storage
- Facilitate local restart on certain failures
- Cluster manager
  - ensures shared FS is only active on one of the hosts at any given time.
  - triggers fencing on node failure before failover.

HADR != Mutual Failover

- Similar
  - Multiple instances, DBs
- But differ in many ways
  - Setup / configuration
  - Prerequisites
  - Failure behaviour
  - Resource Model
  - Monitoring

HA Properties	HADR (Pacemaker)	Mutual Failover (Pacemaker)
Maximum cluster nodes	exactly 2	exactly 2
Quorum	Qdevice	Qdevice
Auxiliary standby	max 2	No
Read-On-Standby	Yes	No
Shared storage	No	Yes
Peer DB setup	Yes	No
Log Shipping	Yes	No
I/O Fencing	No	Yes
Monitor Database	Yes	No
Monitor Mounts	No	Yes
Monitor Instance	Yes	Yes
Monitor Network	Yes	Yes
Monitor Virtual IP	Yes	Yes

Summary

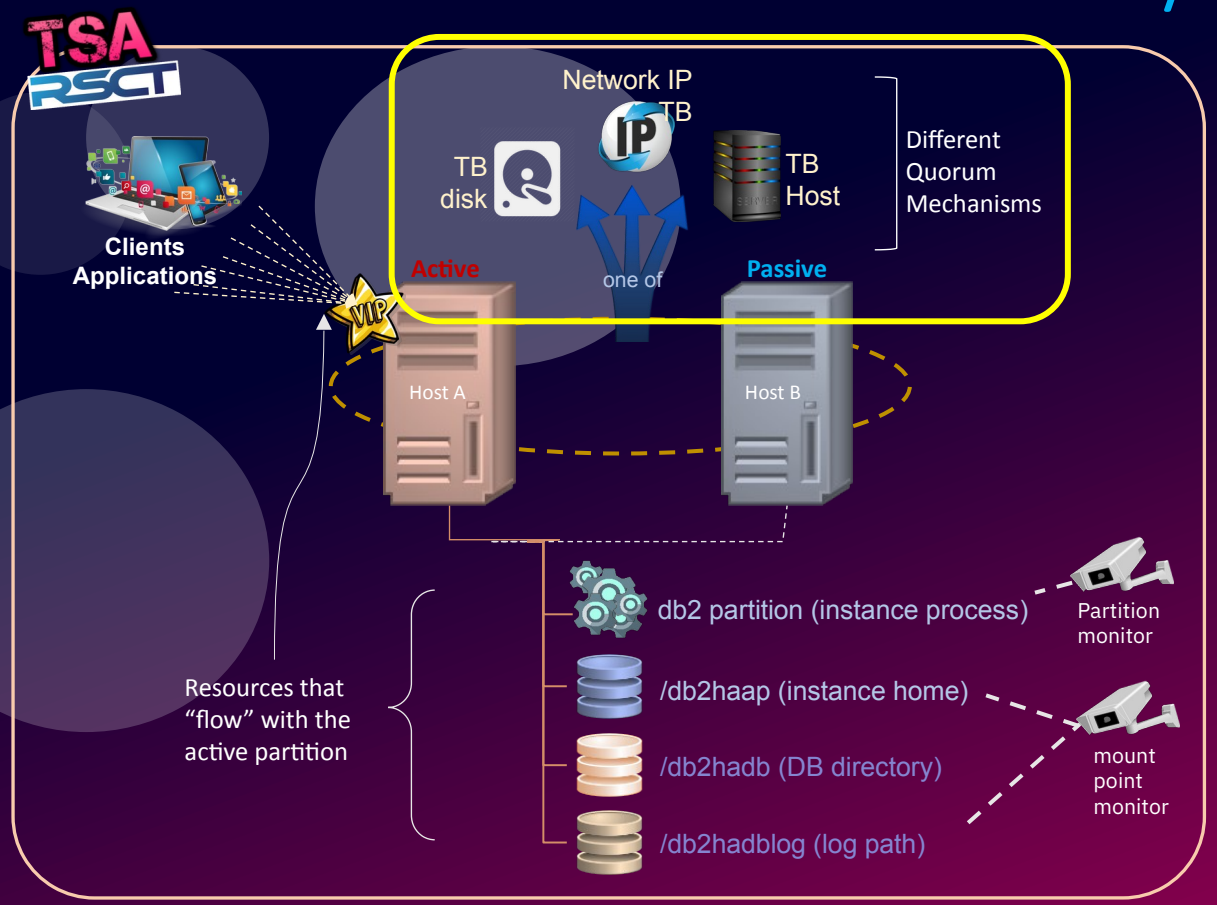
HADR

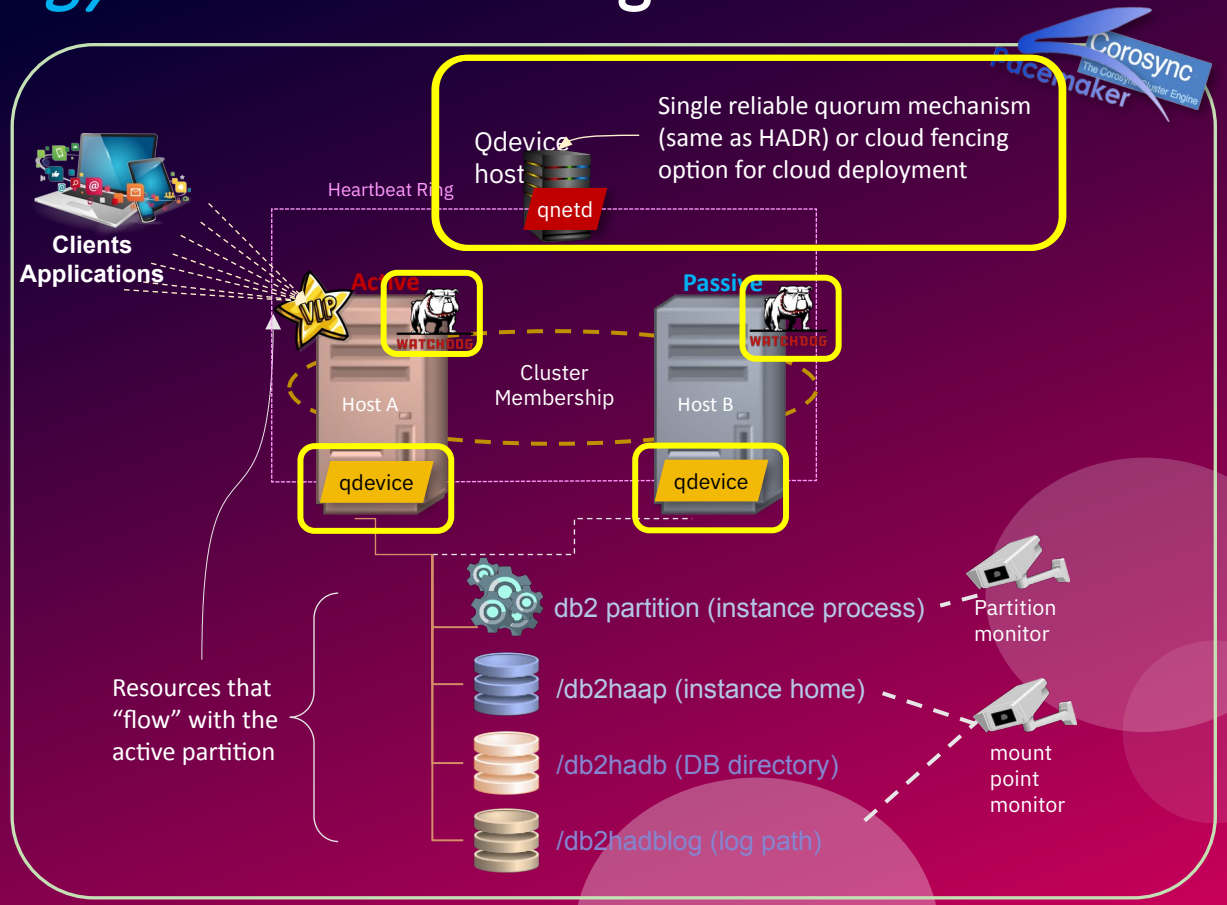
- More granular failover at DB level
- Same instance on both hosts can be active at the same time

Mutual Failover

- All resources to fall under partition, leading to all or nothing failover
- Every resource can only be active on one host

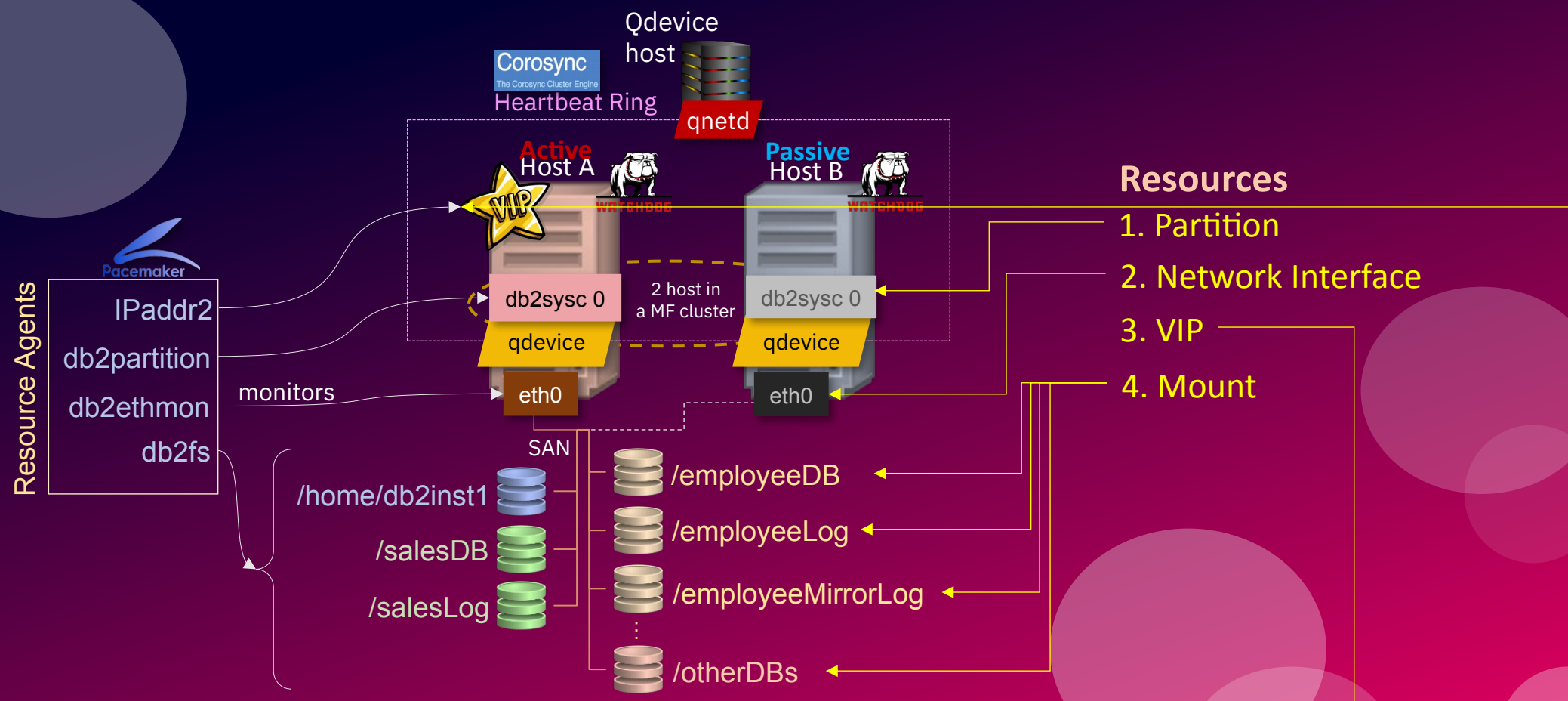
2-node Mutual Failover HA: *Topology Old Vs New* at a glance



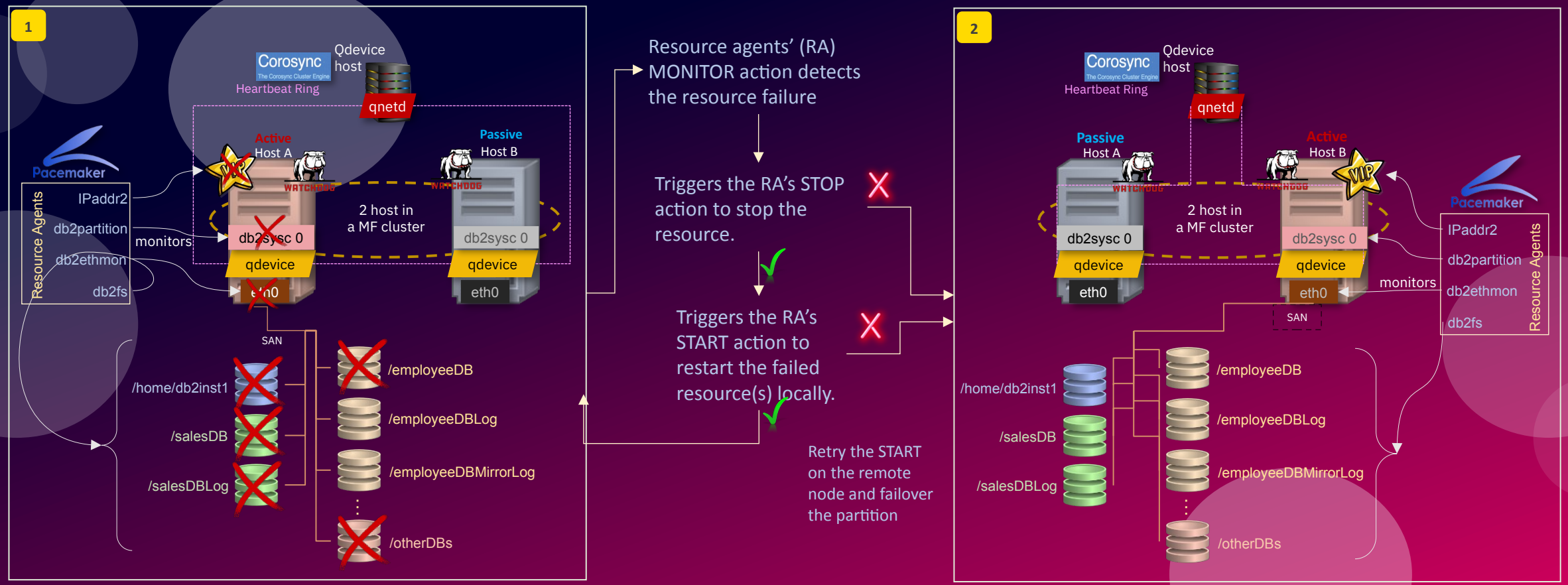


TB disk: requires SCSI 2/3 (not cloud-friendly), Network IP: not reliable, TB host: heavy handed	Quorum	QDevice has lightweight non-H/W requirements, reliable, cloud-ready, multi-clusters, cross-arch.
Rely on RSCT Critical Resource Protection to reboot when a resource failed	Fencing	Leverage OS Software Watchdog to reboot host when a fencing action is required
Engine is integrated, but not every Db2 utility is cluster-aware	Engine Integration	Engine and utilities are cluster-aware such as db2relocatedb
No	Cloud-Readiness	Yes
db2haicu - Text-based menu driven utility	Cluster Management	db2cm – option-based command line utility

Architecture: Overview of Resource Models Components



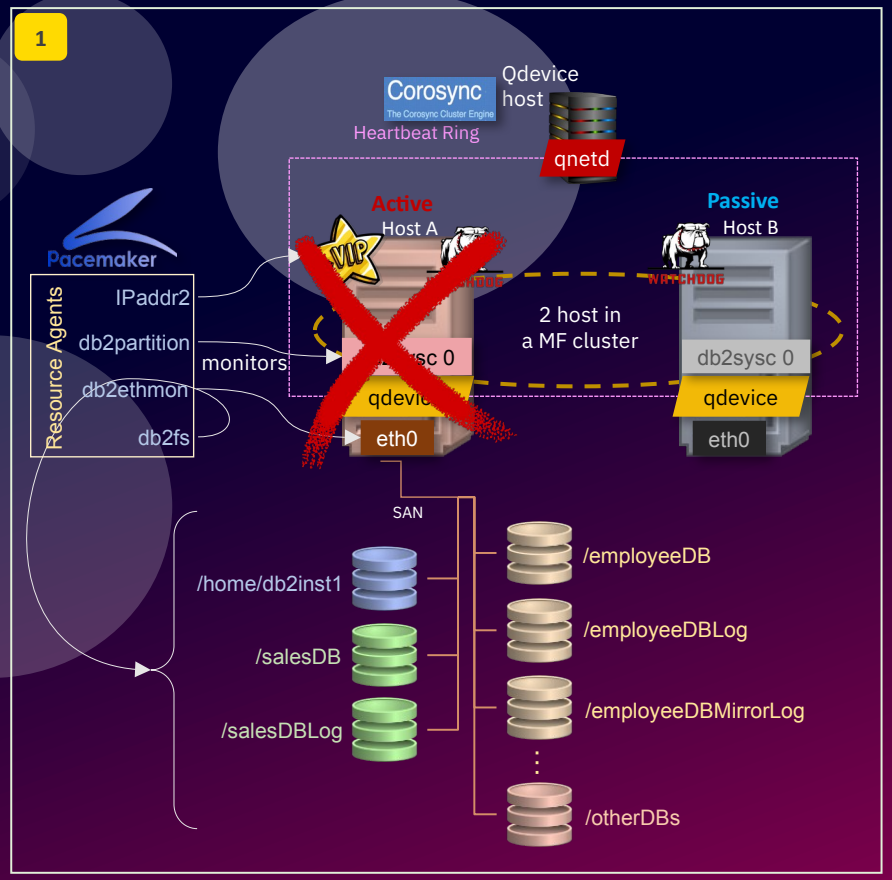
Failure Behaviour: *Resource Failure*

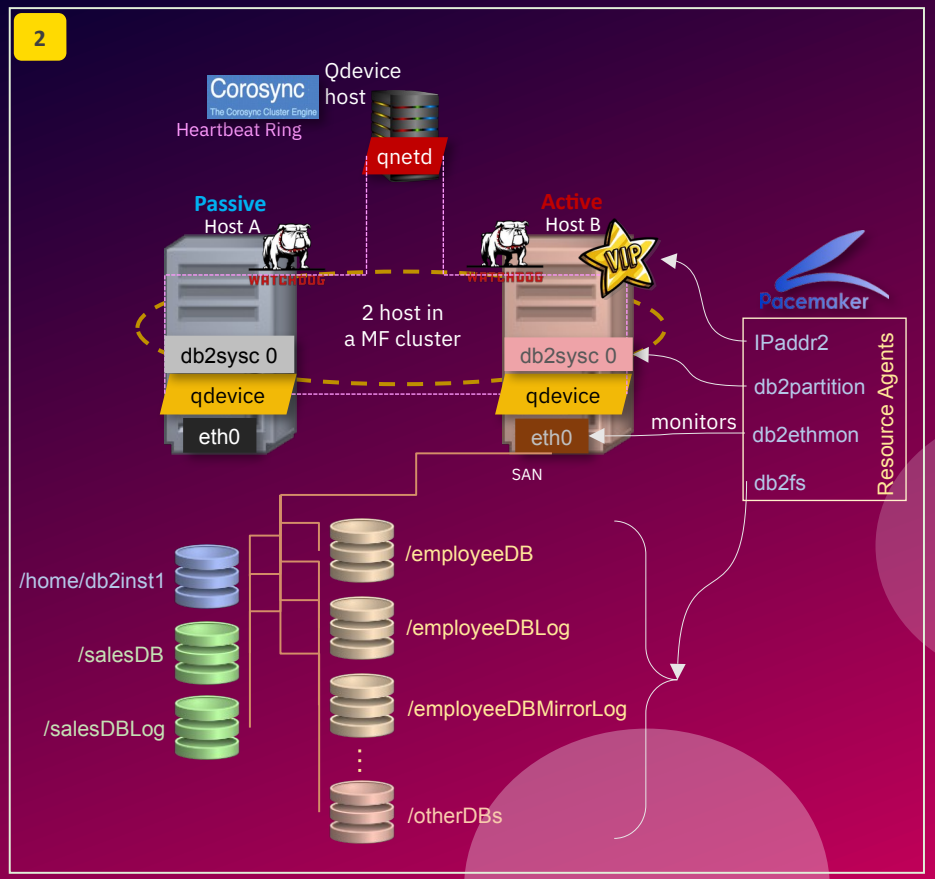


Result

- 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

Recovery Performance compared with TSA

HADR

- 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.*

Cloud Enhancements

1. Enable for both HADR and Mutual Failover HA configurations !

2. Updated fencing agent on both AWS and Azure
   - Updated Fence agent available via the IBM hosted – Market Registration Site (MRS)

Description	Filename	Size	Action
Db2_RHEL_AWS_fence_agents_4.11.0-4.tar.gz	Db2_RHEL_AWS_fence_agents_4.11.0-4.tar.gz	1161961 B	Download
Db2_RHEL_Azure_fence_agents_4.11.0-4.tar.gz	Db2_RHEL_Azure_fence_agents_4.11.0-4.tar.gz	1170586 B	Download
Db2_SLES_AWS_fence_agents_4.7.1-3.tar.gz	Db2_SLES_AWS_fence_agents_4.7.1-3.tar.gz	957324 B	Download
Db2_SLES_Azure_fence_agents_4.9.0.tar.gz	Db2_SLES_Azure_fence_agents_4.9.0.tar.gz	670806 B	Download

3. Automated setup for VIP setup using cloud vendor specific technology
   - AWS: Overlay IP *(Reduce 8+ manual steps to 1 for a single VIP setup)*
   - Azure: Load Balancer *(Reduce 5 manual steps to 1)*

4. Automated setup for alternate quorum (no 3rd host) via cloud vendor fencing
   - AWS fence agent
   - Azure fence agent

HADR

Cloud Enhancement – Automate *AWS VIP (Overlay IP)* setup

End-to-end setup overview

1. Provision EC2 instances and HADR cluster setup
2. Determine the IP address and NIC for Overlay IP
3. Disable source/destination check in both instances
4. Create and associate an AWS policy to IAM role
5. Create an AWS profile and set region
6. Update routing table with OIP to point to primary instance
7. Add Overlay IP to primary node
8. Add OIP resource to resource model
9. Change VIP resource to be managed by Pacemaker

db2cm -list

```
Resource Name    =
db2_db2inst1_db2inst1_CORAL-primary-
OIP
  State                    = Online
  Managed                  = true
  Resource Type            = IP
    Node                   = ip-10-1-15-31
    Ip Address             = 192.168.1.90
  Location                 = ip-10-1-15-31
```

```
db2cm -create -aws {-primaryvip|-standby} <ip address> -rtb <route table id>
      [-profile <profile>] -db <dbname> -instance <instance name>

db2cm -delete -aws -primaryvip|-standby} <ip address> [-profile <profile]
      -db <dbname> -instance <instance name>
```

Full instructions : link

HADR

Cloud Enhancement – Automate *AWS fence agent* setup

End-to-end setup overview

1. Provision VMs, install Db2, and download AWS fence agent

2. Setup HADR cluster

3. Create policy for the instance granting privilege to AWS Centralized VM Control

4. Enable three Pacemaker fencing related config. parameters.

5. Configure fence_aws fencing agent resource in the resource model

Consolidate into new db2cm options

```
db2cm -create -aws -fence
db2cm -delete -aws -fence
```

db2cm -list

```
Resource Name    = fence_db2_aws
  State                  = Online
  Managed                = true
  Resource Type          = Fence Agent
  Current Host           = ip-10-1-15-31

Fencing Information:
  Configured
```

Full instructions : link

HADR

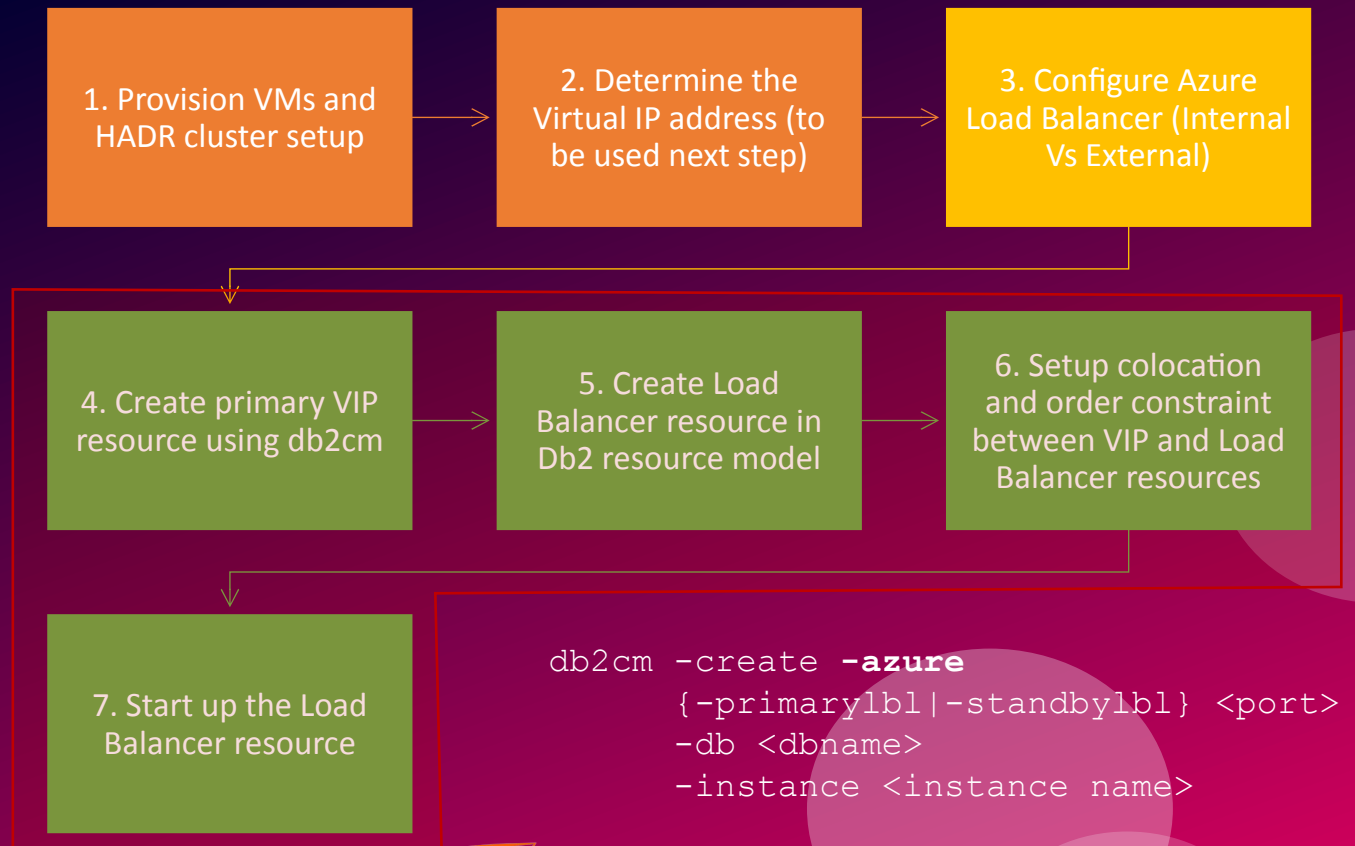
Cloud Enhancement – Automate *Azure VIP Load Balancer* setup

db2cm -list

```
Resource Name     =
db2_db2inst1_db2inst1_CORAL2-primary-
VIP
  State                  = Online
  Managed                = true
  Resource Type          = IP
    Node                 = Host-2
    Ip Address           = 10.0.0.52
  Location               = Host-2

Resource Name     =
db2_db2inst1_db2inst1_CORAL2-primary-
lbl
  State                  = Online
  Managed                = true
  Resource Type          = Load Balancer
    Port                 = 62500
  Location               = Host-2
```

End-to-end setup overview




```
db2cm -create -azure
      {-primarylbl|-standbylbl} <port>
      -db <dbname>
      -instance <instance name>

db2cm -delete -azure
      {-primarylbl|-standbylbl} <port>
      -db <dbname>
      -instance <instance name>
```

Full instructions : link

HADR

Cloud Enhancement – Automate *Azure fence agent* setup

```
db2cm -create -azure -fence
db2cm -delete -azure -fence
```

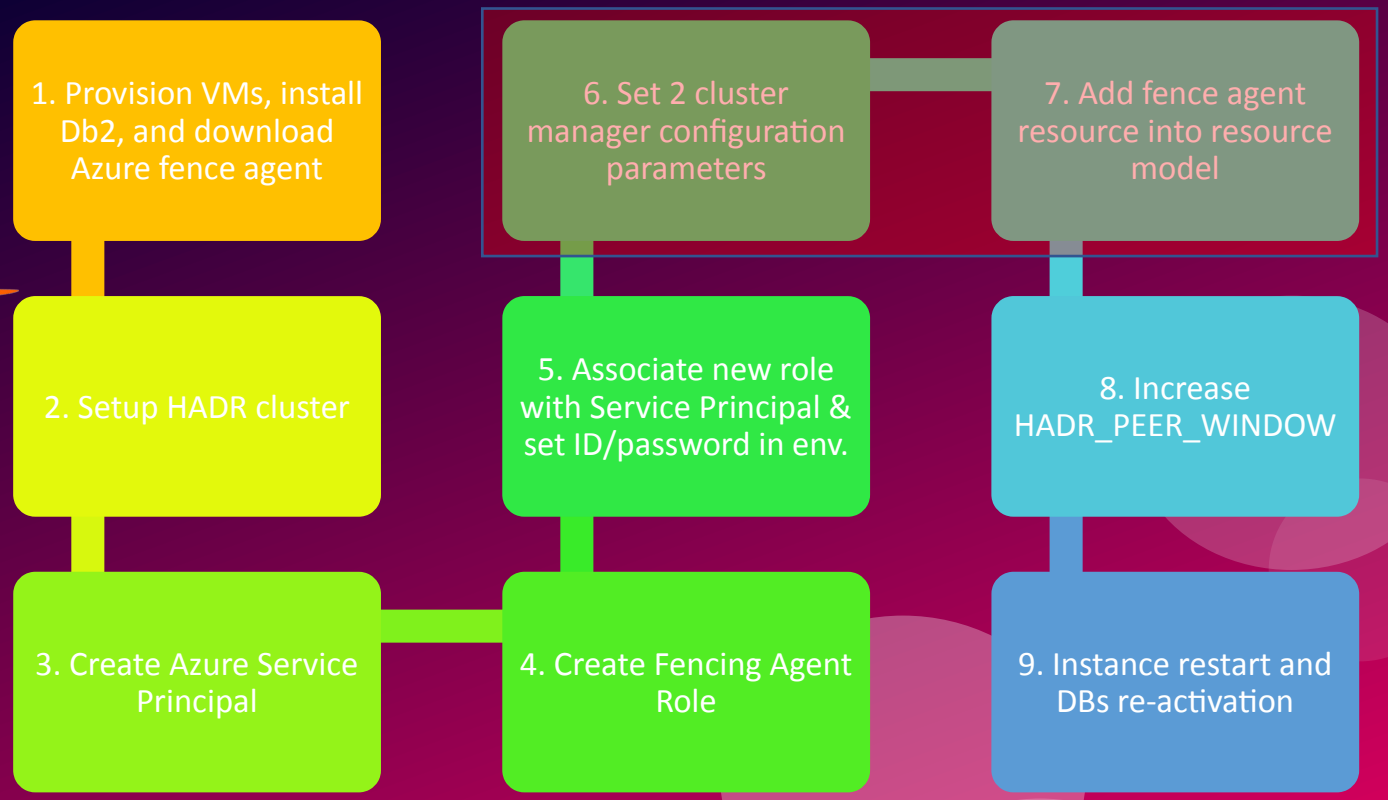
Consolidate into new db2cm options

db2cm -list

```
Resource Name      = fence_db2_azure
  State                = Online
  Managed              = true
  Resource Type        = Fence Agent
  Current Host         = Host-1

Fencing Information:
  Configured
```

End-to-end setup overview



Full instructions : link

Improved Up & Running with Pacemaker: Silent + GUI Install

- 11.5.6.0 supports Pacemaker bundled and auto-install with command line – db2_install
- 11.5.8.0 completes the story with support of *Silent* and *GUI* install

Silent Install command: `db2setup -r <response-file>`

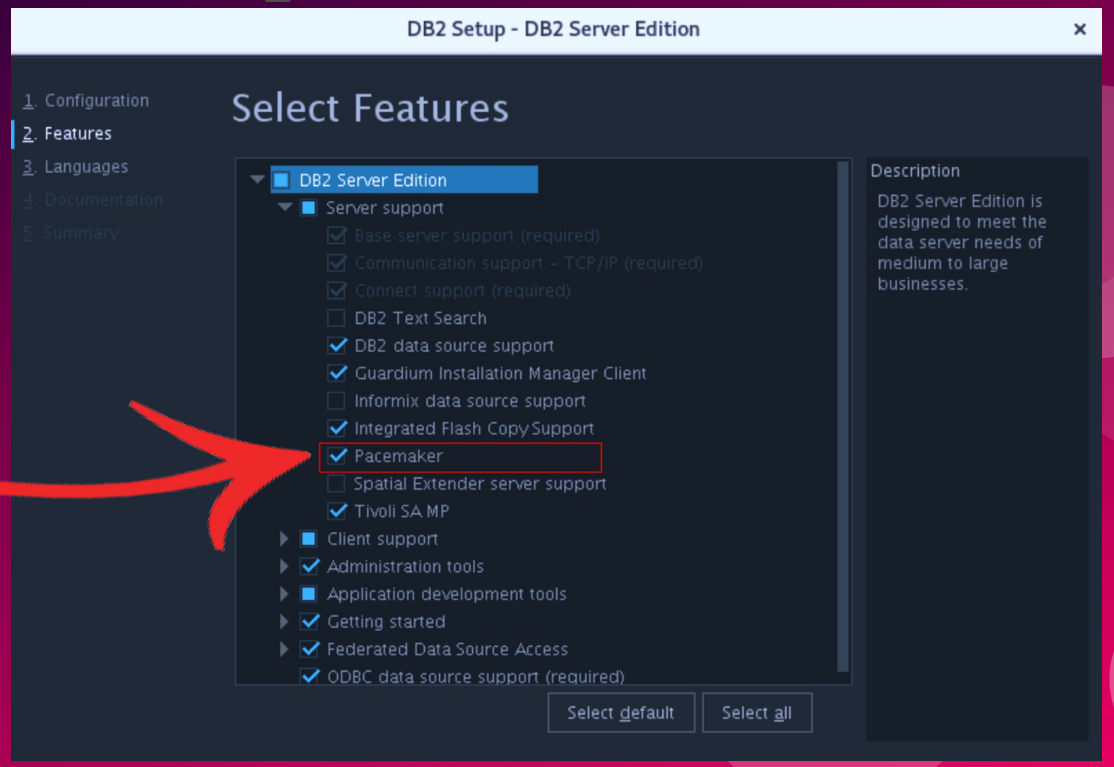
Snippet of response file

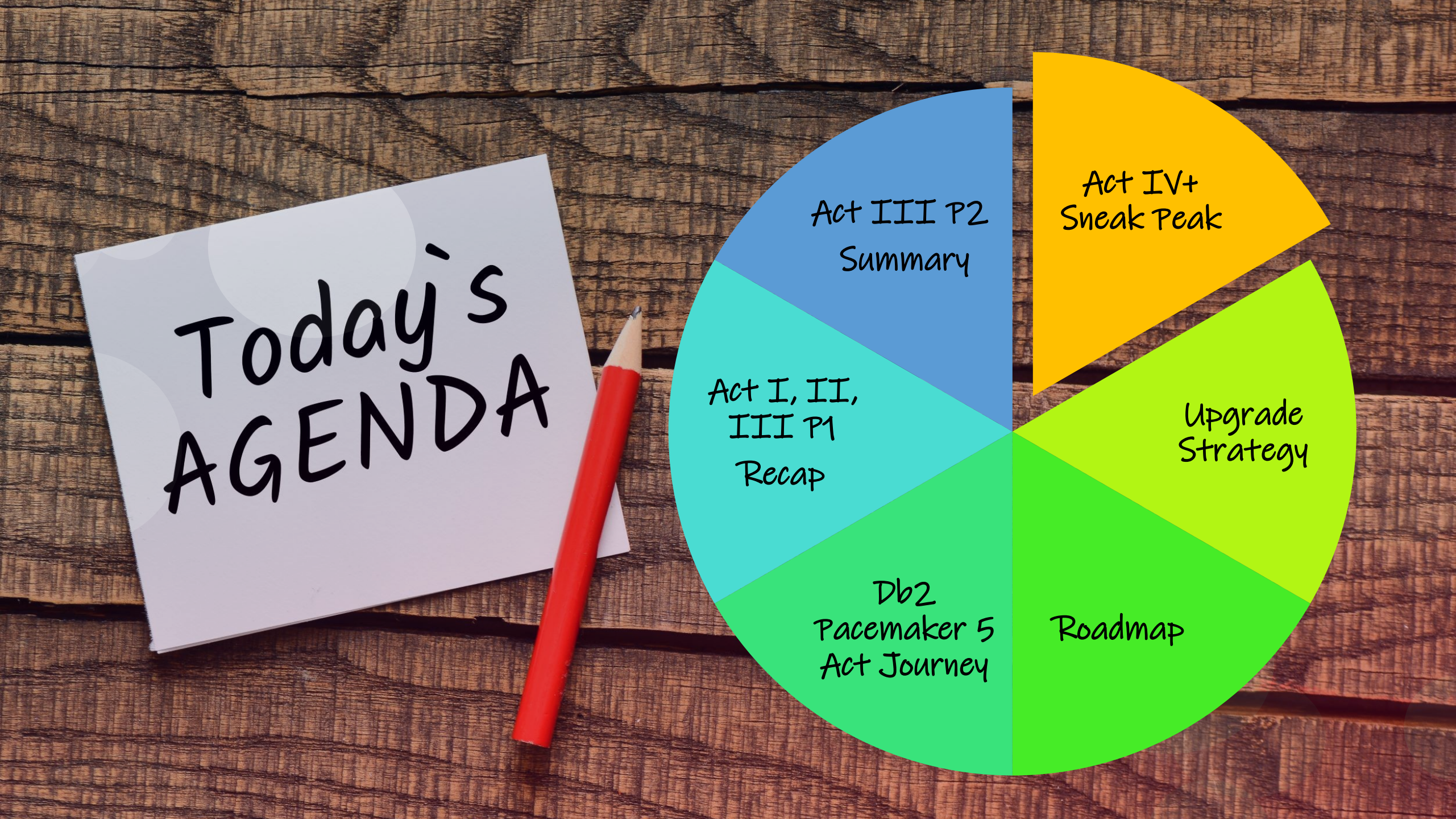
```
PROD = DB2_SERVER_EDITION
FILE = /opt/ibm/db2/install_dir
LIC_AGREEMENT = ACCEPT
INSTALL_TYPE = TYPICAL
INSTALL_TSAMP = <YES|NO>
INSTALL_PCMK = <YES|NO>
INSTANCE = DB2_INST
DB2_INST.NAME = db2inst1
DB2_INST.GROUP_NAME = db2iadm1
DB2_INST.PASSWORD = password1234
DB2_INST.TYPE = ese
DB2_INST.START_DURING_INSTALL = YES
DB2_INST.FENCED_USERNAME = db2sdfe1
DB2_INST.FENCED_GROUP_NAME = db2fsdm1
DB2_INST.FENCED_PASSWORD = password123
```

optional

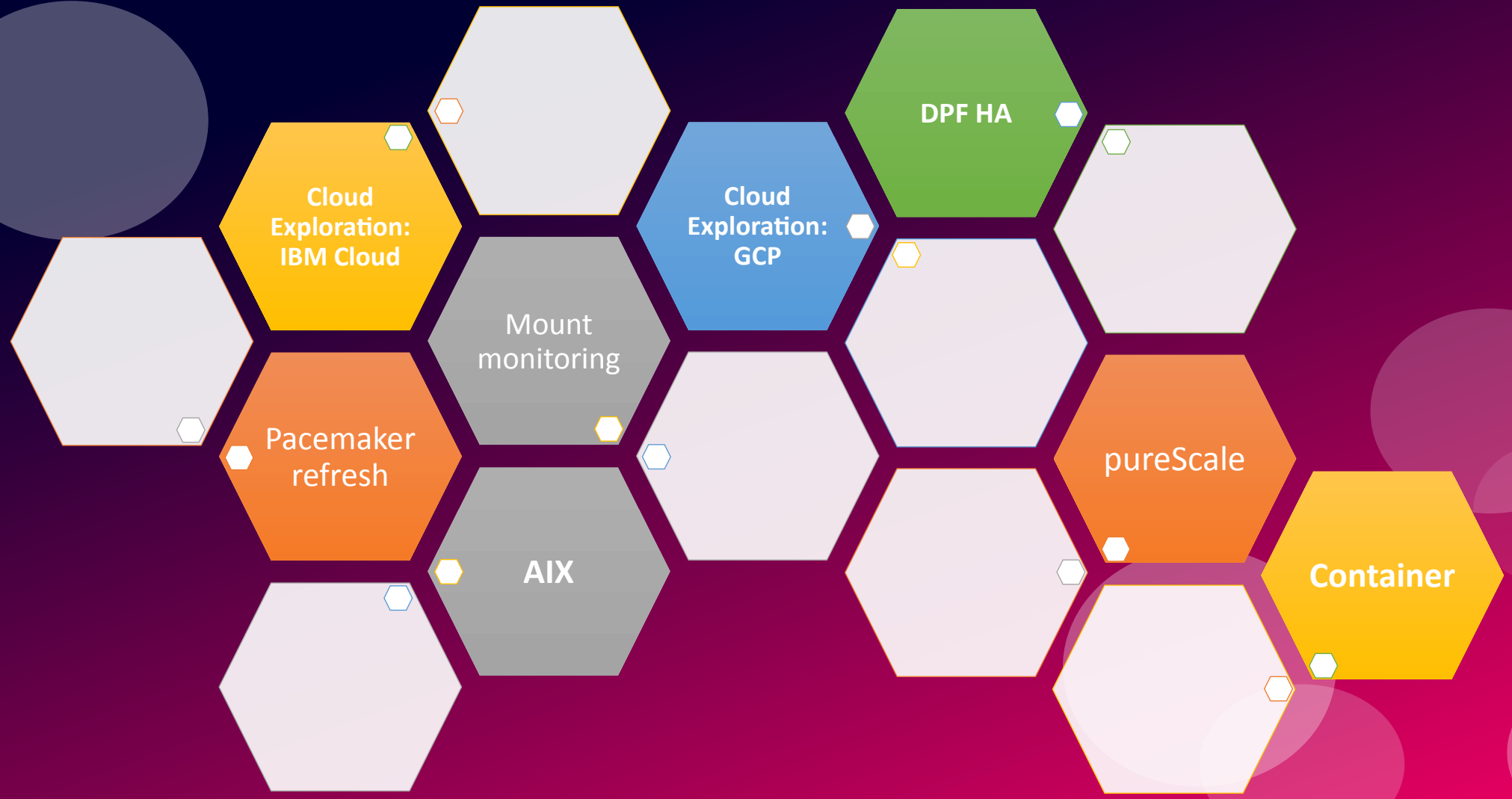
Note: The two keywords do ***not*** depend on each other.

GUI Install command: `db2setup`



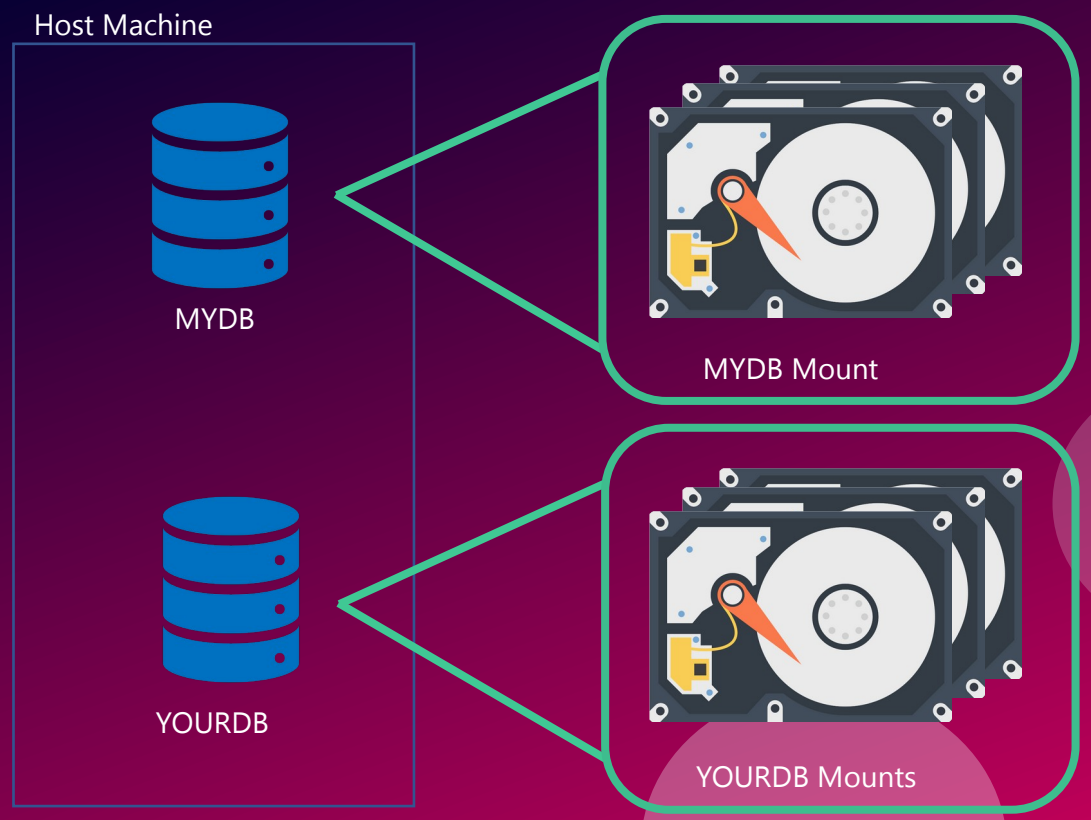

Today`s AGENDA
Act III P2 Summary
Act IV+ Sneak Peak
Act I, II, III P1 Recap
Upgrade Strategy
Db2 Pacemaker 5 Act Journey
Roadmap

Upcoming Key features …



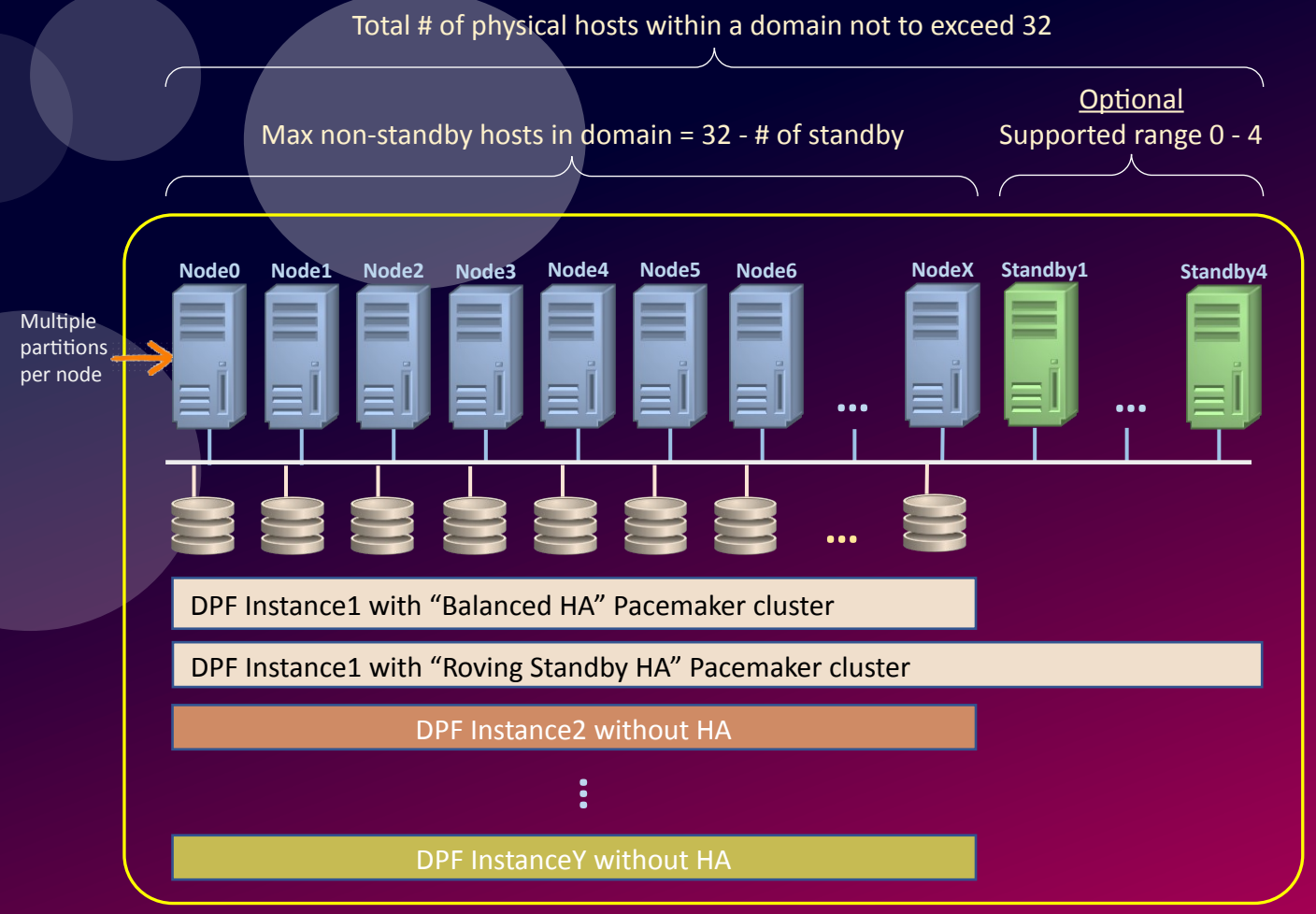
Mount Automation in HADR with Pacemaker

- Make filesystems highly available
- Adds order constraint between the database and its associated filesystems.
- Ensures the database filesystems are operational before a database is activated.
- Attempt to automatically bring filesystems back online in failure scenarios.
- Used in various topologies.



Feature inclusion in future Db2 release subjects to change without notice.

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:

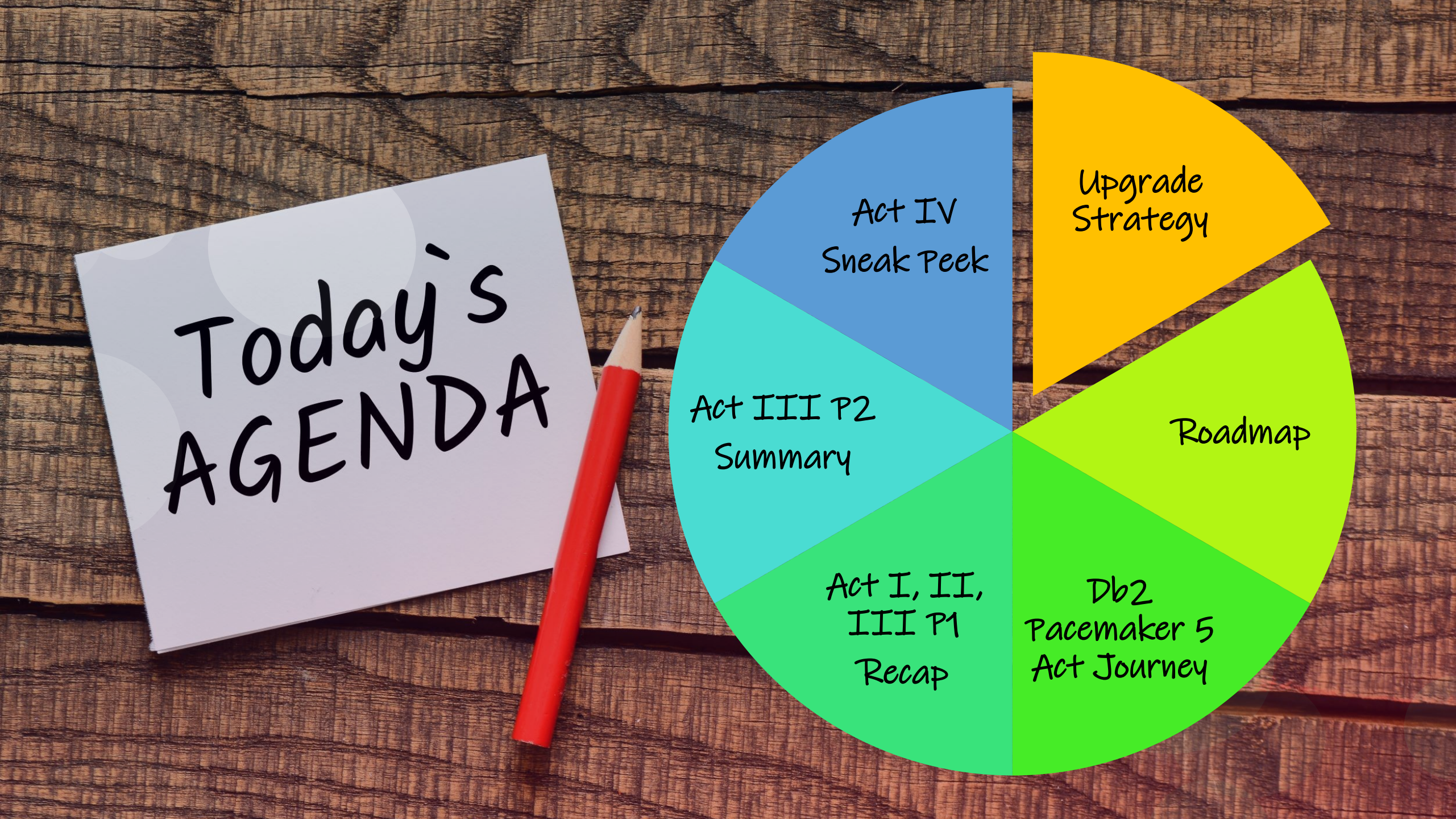
- Higher number of partitions can explore deploying more partitions per host than 8 with proper H/W

pureScale … a teaser

expect

- Cloud-Ready !!!!!
- New & Simplified Resource Model
- Different quorum mechanism (fewer shared disk requirement)
- Db2-optimized node-liveliness test
- More accurate RDMA network liveliness test
- Built-in RDMA network performance evaluation and aggregate history
- Smarter unified cluster management utility interface
- Reduced dependency in support infrastructure
- … and many others

Feature inclusion in future Db2 release subjects to change without notice.


Today`s AGENDA
Act IV Sneak Peek
Upgrade Strategy
Act III P2 Summary
Roadmap
Act I, II, III P1 Recap
Db2 Pacemaker 5 Act Journey

HADR: Conversion from TSA to Pacemaker

Upgrade/Update Strategy from Pre-11.5.8.0 TSA HADR cluster …

- Move to 11.5.8.0 with TSA FIRST, then convert to Pacemaker
- No direct export from TSA and import into Pacemaker.

Backup existing TSA config

- Optional (in case for undo)
- db2haicu -o <backup.xml>

Remove TSA cluster

- Delete all resources, domain, software on both hosts

~~**Install Pacemaker cluster software**~~

- ~~Download from IBM site~~
- ~~Apply to both hosts~~

No longer needed if you are on 11.5.6.0 and up due to integrated bundling and installation support

Create Pacemaker cluster & resources

- Use new cluster management utility, db2cm
- Follow documentation

Validate the cluster

It's online!

*except when VIP is used

V11.5.4.0

*No production use

V11.5.5.0+

*Production for on-prem and non-containerized cloud

Follow the instructions in [Upgrading Db2 servers in a TSA automated HADR environment]

Mutual Failover: Conversion from TSA to Pacemaker

Upgrade/Update Strategy from Pre-11.5.8.0 TSA Mutual Failover cluster …

- Move to 11.5.8.0 with TSA FIRST, then convert to Pacemaker
- No direct export from TSA and import into Pacemaker.

Upgrade to 11.5.8.0 using TSA

Backup TSA cluster
- Optional
- db2haicu -o <backup.xml>

Remove TSA cluster
- Delete all resources, domains, software on both hosts.
- db2haicu -delete

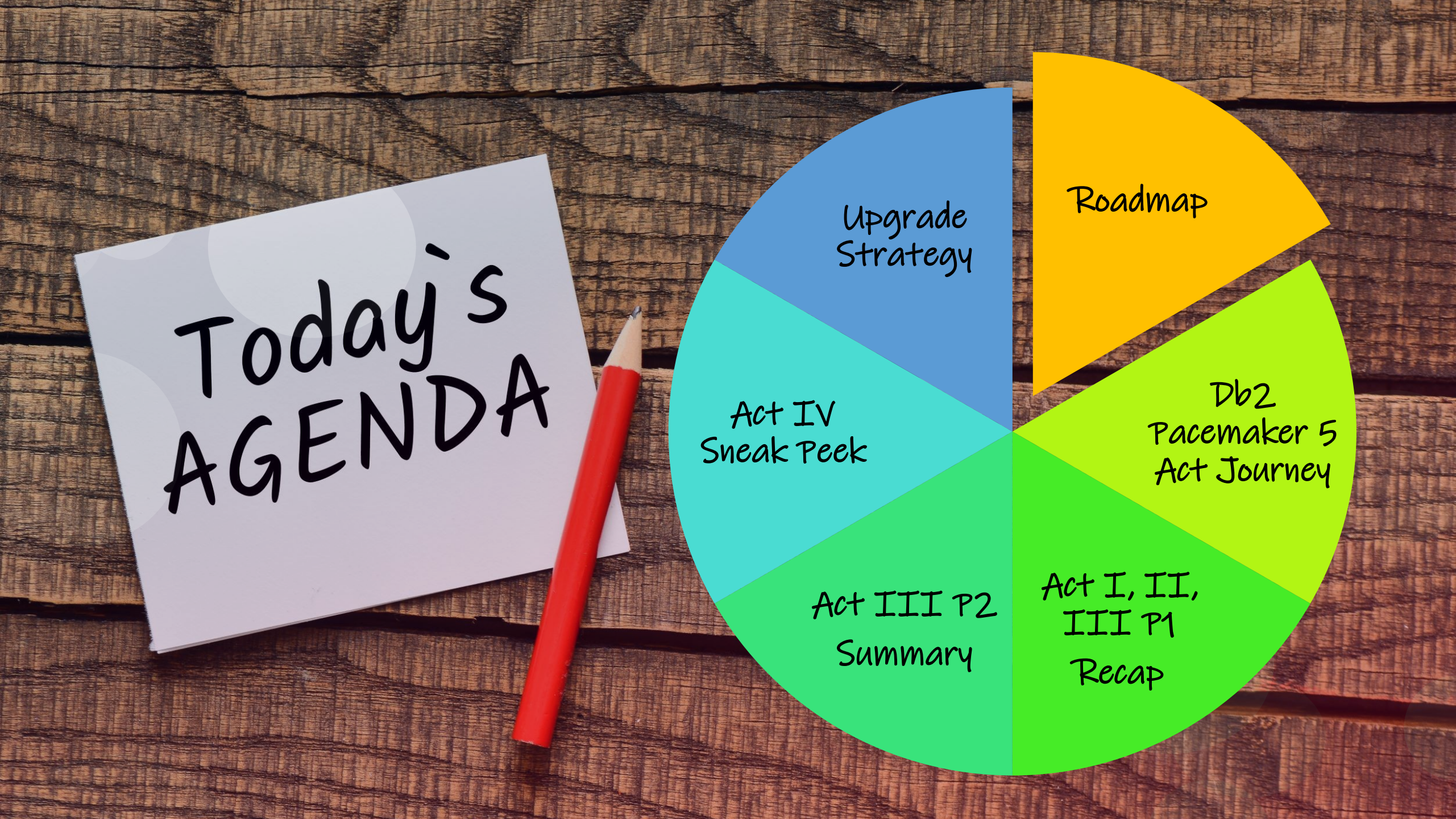
Create Pacemaker cluster & resources
- Use db2cm

Validate the cluster

Take Note

11.5.8.0 MF with Pacemaker will serve as the base for any future upgrade from pre-Next major release as TSA will no longer be supported in VNext. All previous releases using MF must convert to 11.5.8.0 with Pacemaker before upgrade to next major release.

Follow the instructions in [Upgrading Db2 servers in a TSA automated HADR environment]


Today`s AGENDA
Upgrade Strategy
Roadmap
Act IV Sneak Peek
Db2 Pacemaker 5 Act Journey
Act III P2 Summary
Act I, II, III P1 Recap

Roadmap w.r.t. Pacemaker from feature perspective

V11.5.4.0 Technical Preview

- Cluster manager-aware integrated Db2 commands
- Integrated data collection via db2support
- Multiple instances & databases support
- New cluster manager config utility – db2cm
- Enhanced quorum support with QDevice
- RHEL 8.1, SLES 15 SP1 support on Intel and Linux on IBM Z
- Validated on AWS with RHEL 8.1

V11.5.5.0

Technical Preview → GA

- Multiple Standby Support
- Fast redeployment via import & export support
- Two node support with fencing on AWS
- Newer Pacemaker version

V11.5.6.0

- Integrated bundling and install of Pacemaker stack
- Customized configurations on Azure
- Enhanced Network Resiliency
- Advance HADR DB hang detection (Linux)
- Expanded distro levels support
- Enhanced PD

V11.5.7.0

- POWER Linux Support
- Advance HADR DB hang detection (AIX)
- Pacemaker Refresh

V11.5.8.0

- 2-node Mutual Failover HA
- Silent + GUI Integrated Install
- Simplified setup on cloud
- Pacemaker Refresh

V11.5.9.0 Tentative

- Pacemaker Refresh
- Expanded distro levels support
- Customized config on GCP

VNEXT Tentative

- pureScale
- DPF HA
- Enhanced db2cm Security Model
- Mount Monitoring on HADR
- Expanded distro levels support
- Customized config on IBM Cloud

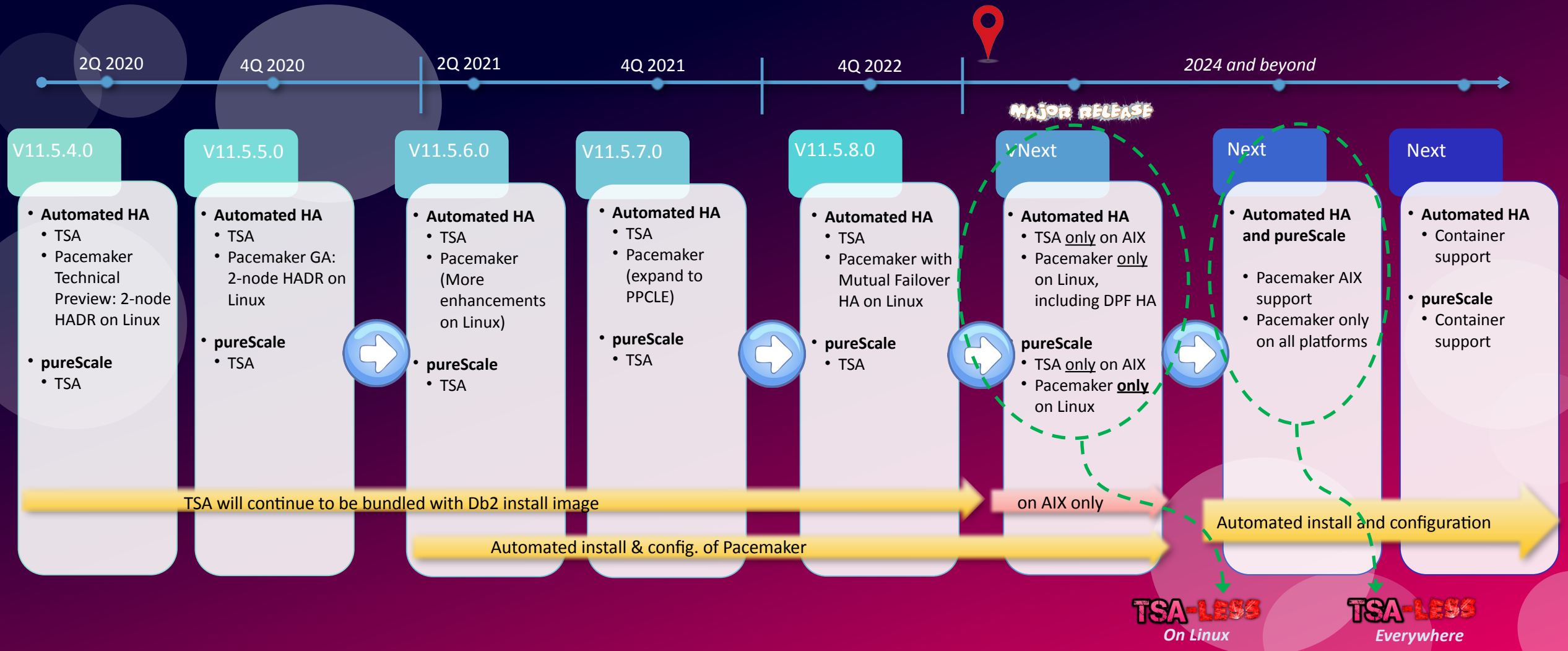
Future GAs Tentative

- AIX support
- Container Support
- Other Aha Ideas

Linux → Linux + AIX

Note: Roadmap and content subjected to change

Overall Roadmap



Note: Roadmap subjected to change

41