IDUG VIRTUAL

2021 MidWest Db2 Virtual Seminar

IDUG

Pacemaker for Db2 LUW: The Foreword, The Preface, The Introduction

Hao Qi, IBM

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

Agenda

The Foreword

- "Why"

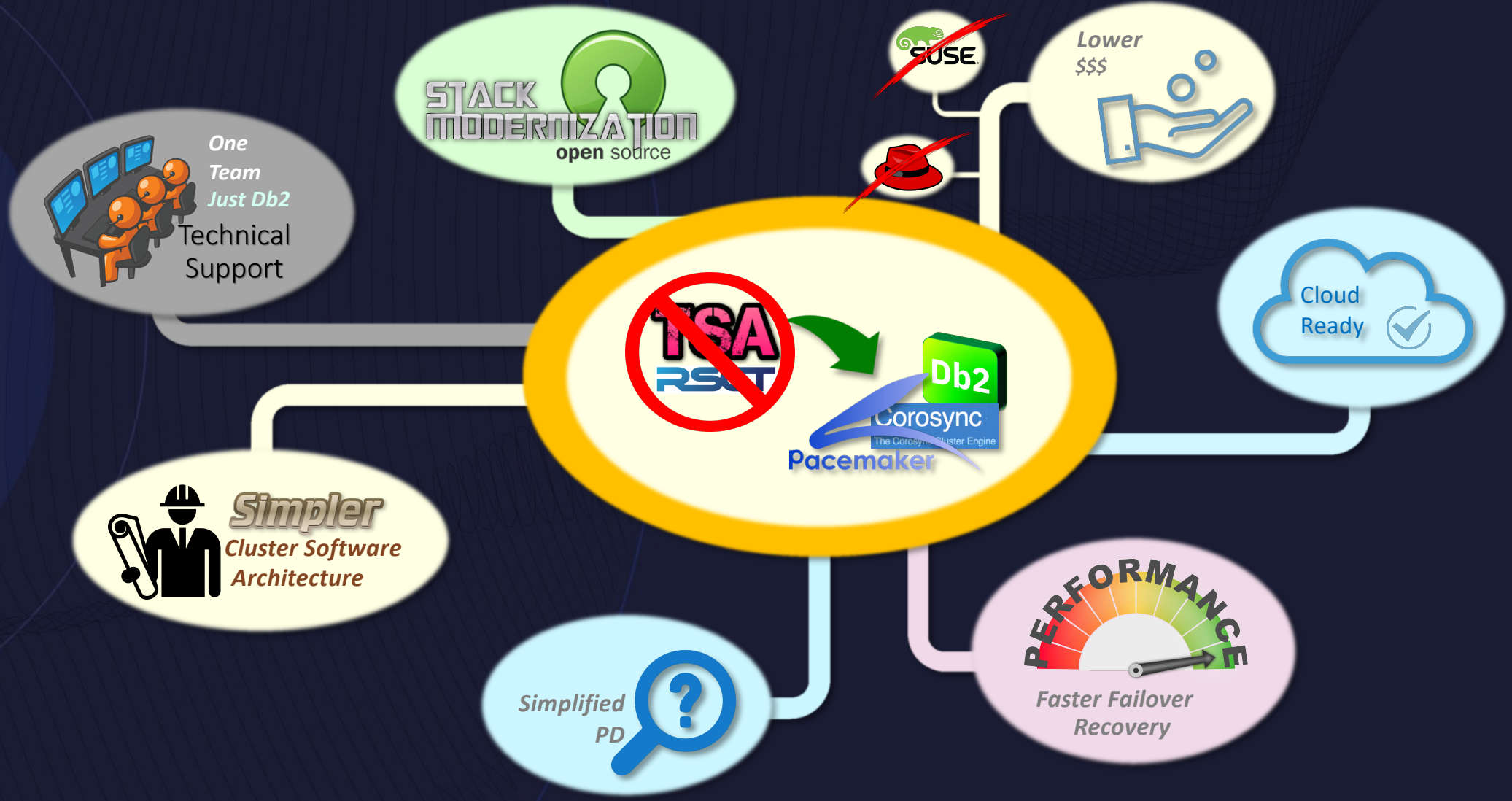
The Preface

- "How"

The Introduction

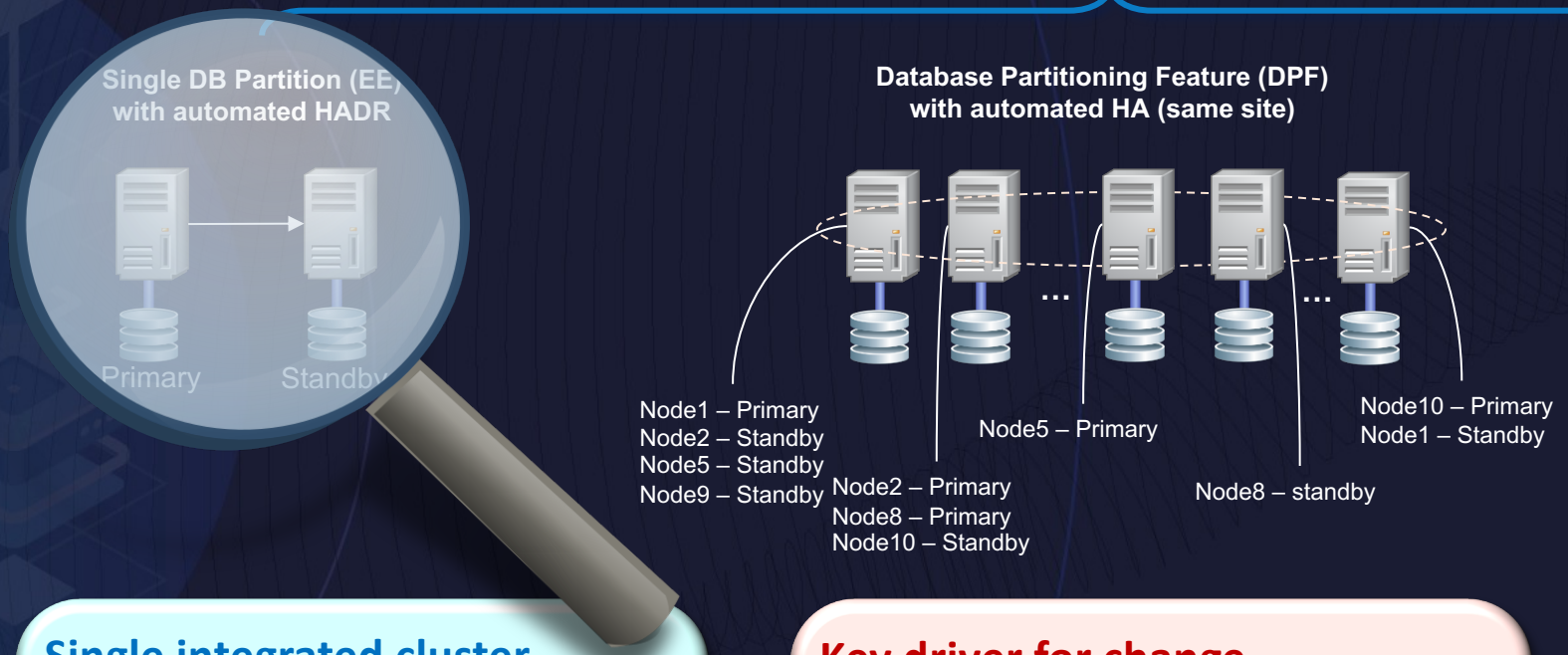
- Many "Whats"

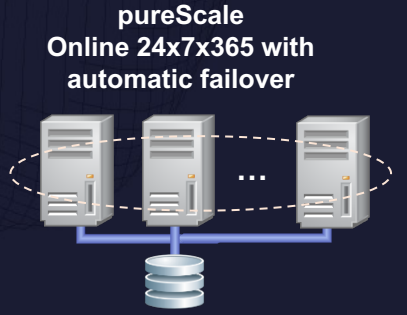
The list of “Why”



The need for a cloud-ready, enterprise-ready Db2 integrated cluster manager

TSA





Single integrated cluster manager

- For all supported platforms
- For all deployment models

Key driver for change

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

Dev Plan

- *Concurrent development in both HADR and pureScale*
- *Quicker release schedule with HADR*

Stack Modernization

Proprietary

Tivoli. software + RSCT

Pacemaker + Corosync The Corosync Cluster Engine

Open Source

Why?

Pacemaker

Corosync The Corosync Cluster Engine

- 17+ 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

Cost savings with Db2 Pacemaker Vs other Pacemaker providers – 2 flavours

1

	Distro-supported Pacemaker	Db2-supported Pacemaker
Per cluster node cost	From ~$400 to ~$800	FREE
Per year support cost	From ~$1200 to ~$2000	FREE

2

	Distro-supported Pacemaker	Db2-supported Pacemaker
What “extras” may be needed to make the Db2 HADR solution work ?	Different distros poses different limitations (max # of nodes, etc.)	We validate with our existing solution – no surprises, no hidden cost
Architecture / platform support	RHEL: Intel and POWER, No IBM Z SLES: Intel, POWER, and Z	RHEL & SLES on Intel and IBM Z AIX is on the roadmap
Time to support the latest greatest HA features from upstream open-source	Vary by distros. Typically, one full release, if not more, lower than up-stream	Adopts the latest greatest from upstream Goal: move up as often as we can
Support turnaround time	Slower	Faster

New and more flexible support statement

As Non-Db2 Integrated solution

Pacemaker + dependent s/w provided by

redhat
High Availability Add-On

ClusterLabs
Open Source

SUSE
SUSE Linux Enterprise High Availability Extension

HA with Pacemaker Corosync
The Corosync Cluster Engine

As Db2 Integrated solution

Pacemaker + dependent s/w provided by

100% CERTIFIED Db2

Cluster Manager related

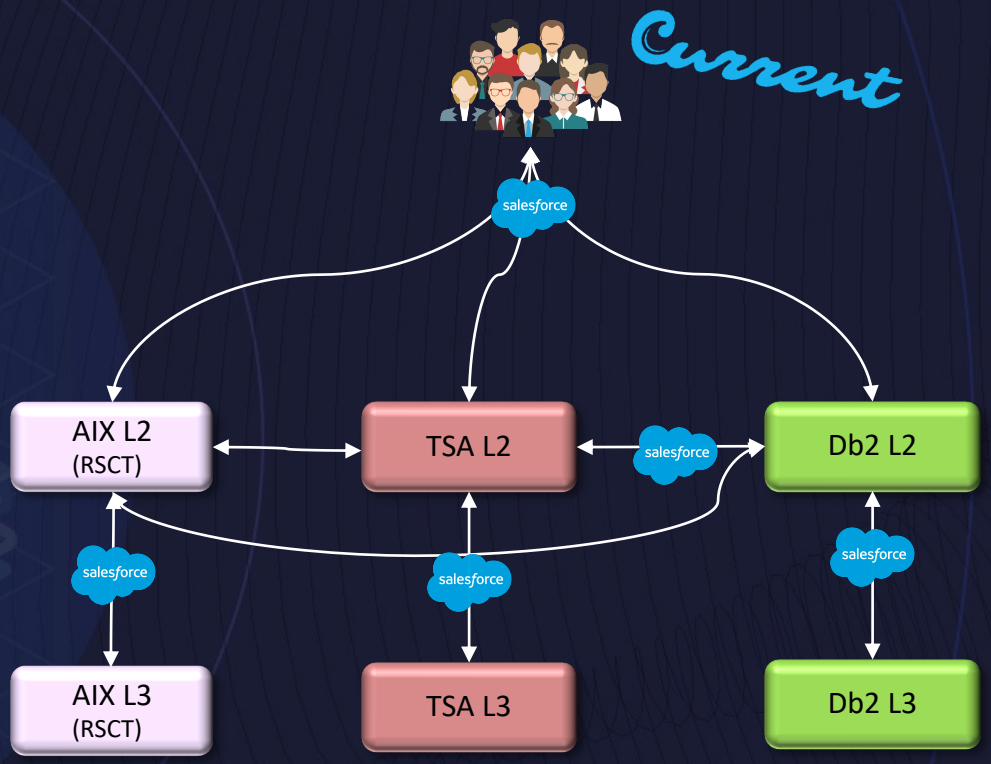
receive support from

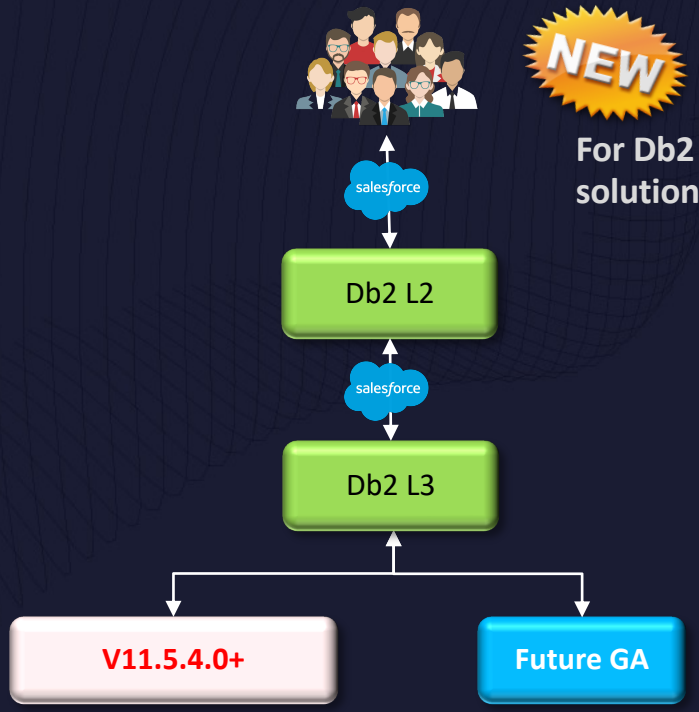
SUPPLIERS

Pacemaker

Db2

Db2 Support Flow – Old Vs New





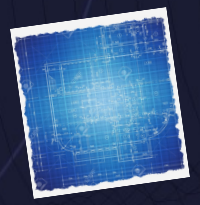
Support Coverage

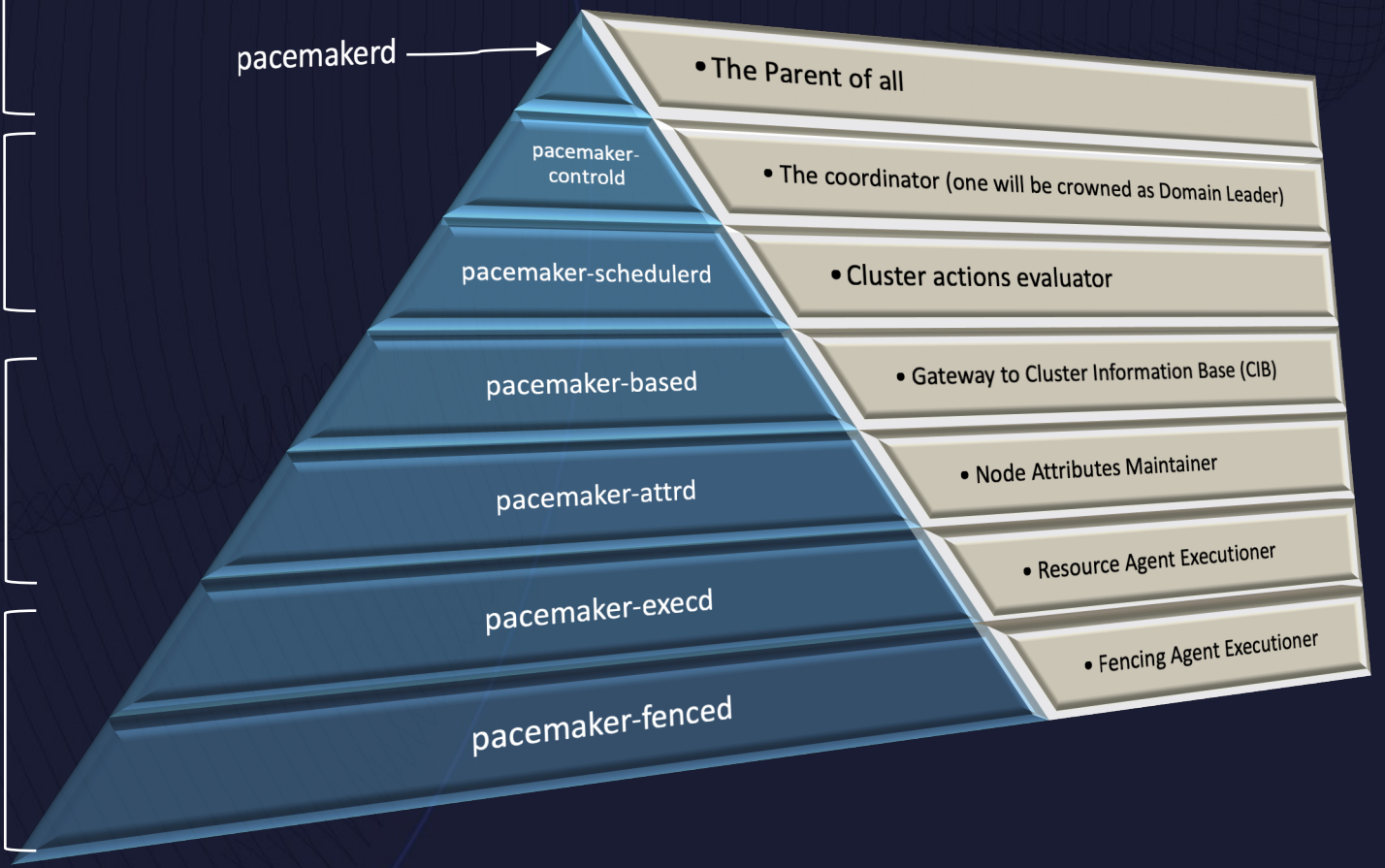
Db2 provides Pacemaker support:

- when the Db2 provided Pacemaker stack is used as its integration cluster manager solution and the Db2 resource model are setup using the db2cm utility provided at the same release level unless otherwise instructed by Db2. Support of any non-Db2 provided Pacemaker version is the responsibility of the appropriate Pacemaker provider, not Db2.
- on functionalities used by the integrated solution only where the functionalities are setup by db2cm utility unless otherwise instructed by Db2. Support on other features available in the Db2 provided Pacemaker stack but not used by the integration solution are not included.

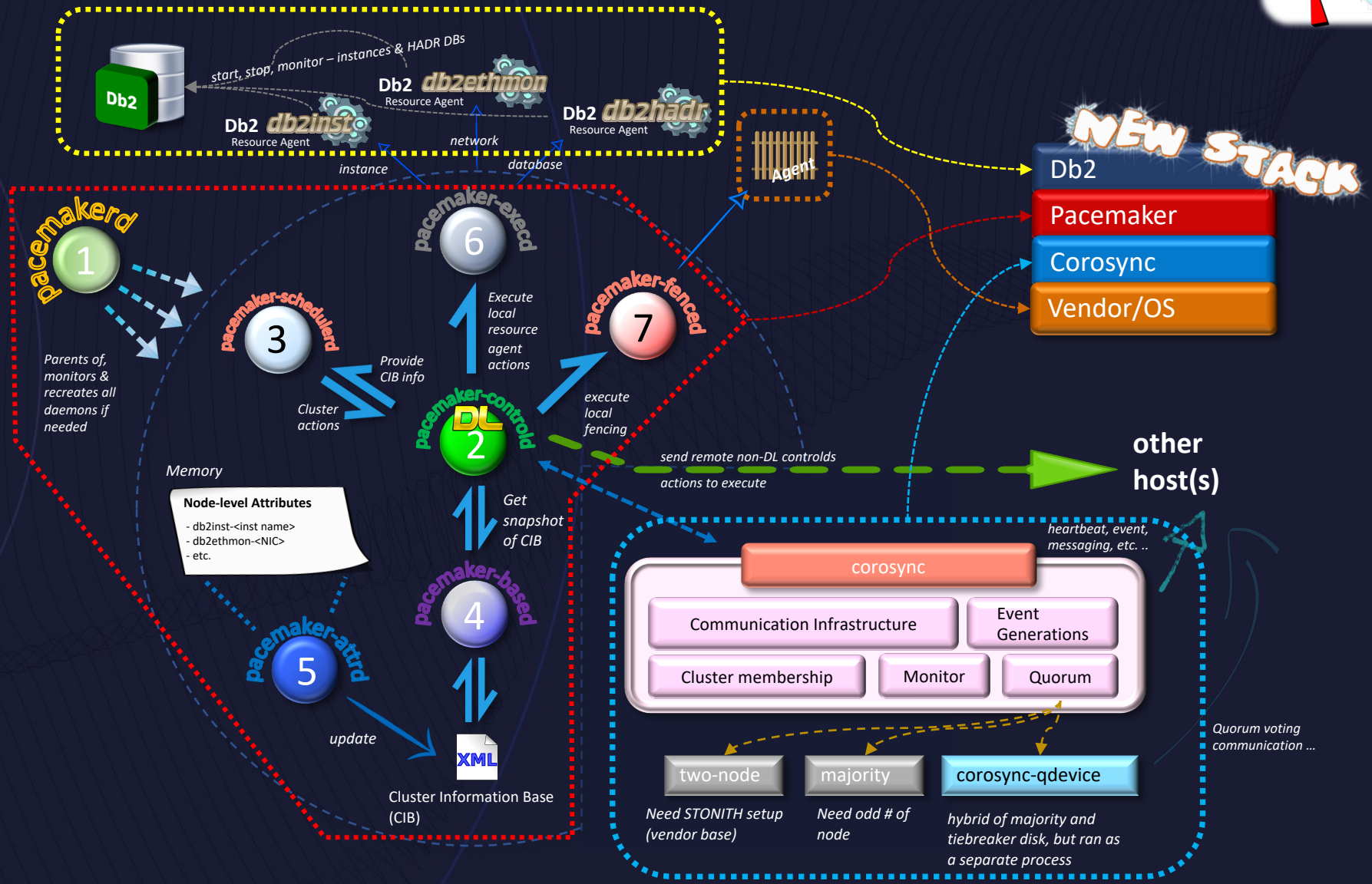
Pacemaker Process Model Overview

```
root      15048     1  0 May15 ?   00:00:45 /usr/sbin/pacemakerd -f
haclust+ 15059 15048  0 May15 ?   00:00:07 /usr/libexec/pacemaker/pacemaker-controld
haclust+ 15054 15048  0 May15 ?   00:00:56 /usr/libexec/pacemaker/pacemaker-based
haclust+ 15057 15048  0 May15 ?   00:01:38 /usr/libexec/pacemaker/pacemaker-attrd
haclust+ 15058 15048  0 May15 ?   00:00:05 /usr/libexec/pacemaker/pacemaker-schedulerd
root      15055 15048  0 May15 ?   00:00:06 /usr/libexec/pacemaker/pacemaker-fenced
root      15056 15048  0 May15 ?   00:00:33 /usr/libexec/pacemaker/pacemaker-execd
```



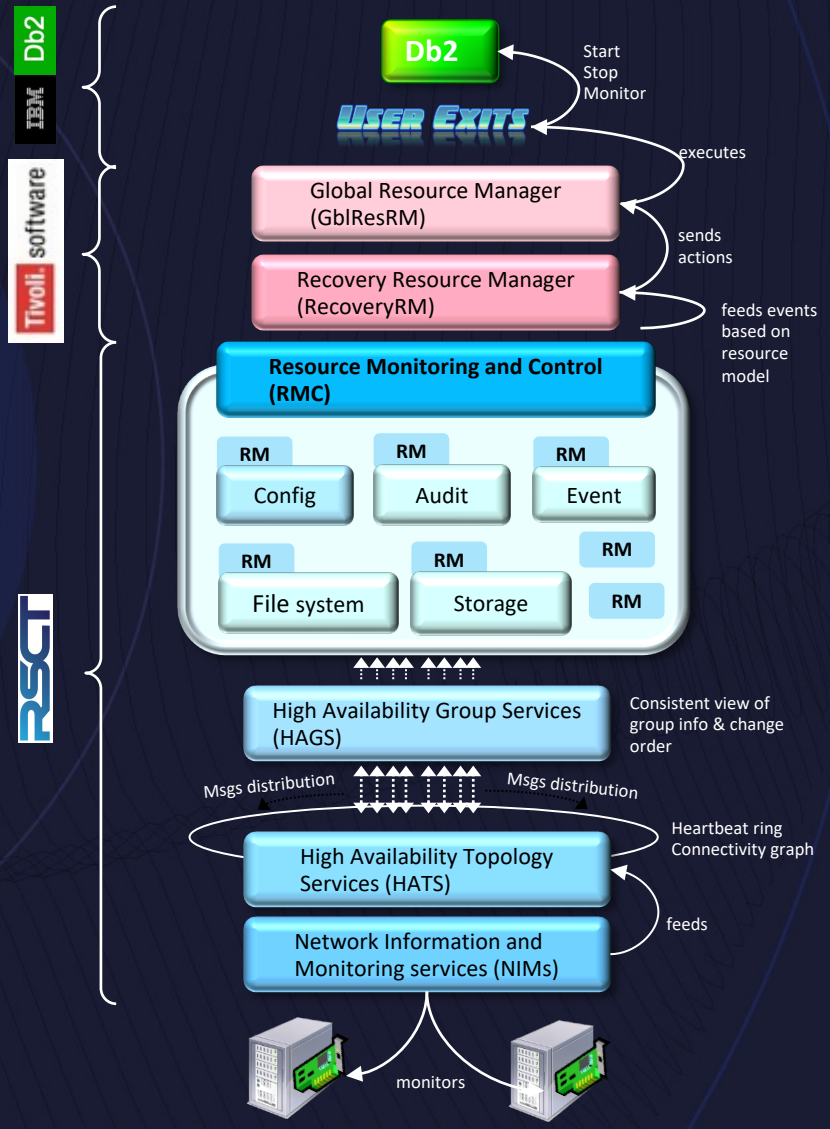


Components Interactions in the “Stack”

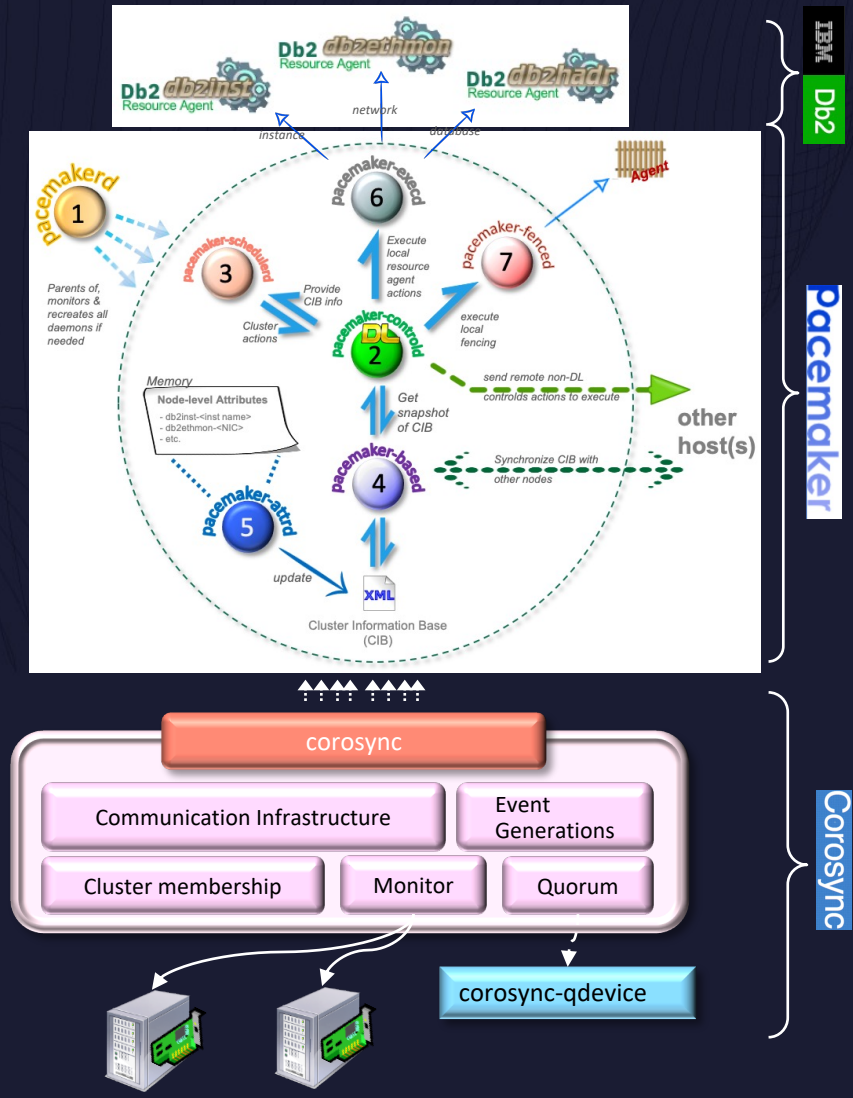


Current Stack vs New Stack

Current HADR Stack – Db2/TSA/RSCT



New HADR Stack – Db2/Pacemaker/ Corosync



TSA Vs Pacemaker – diagnostic files comparison

Tivoli software

Global Resource Manager (GblResRM)

Recovery Resource Manager (RecoveryRM)

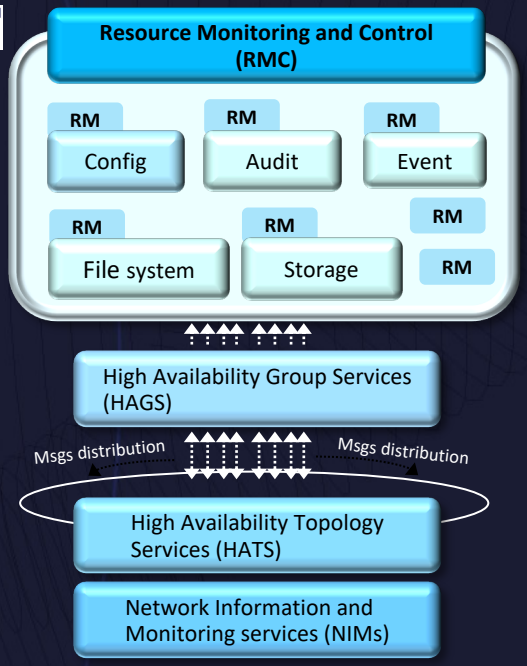
Find RecRM master, then format the trace:

- `lssrc -ls IBM.RecoveryRM | grep -i master`
- `cd /var/ct/<domain_name>/log/mc/IBM.RecoveryRM`
- `rpttr -odtic trace_summary`

Others:

- TSA spool trace, syslog

RSCT



For configRM activities such as fencing, find the leader, then format the trace:

- `lssrc -ls IBM.ConfigRM | grep -i leader`
- `cd /var/ct/IW/log/mc/IBM.RecoveryRM`
- `rpttr -odtic trace_summary >>`

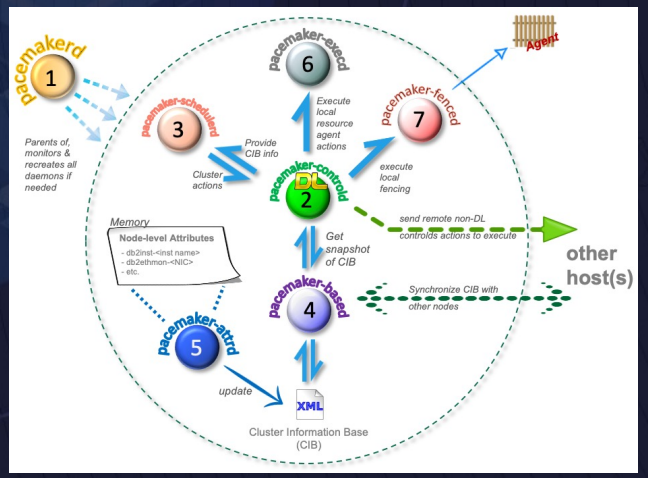
/var/ct/<domain> Vs /var/ct/IW for different RMs

- global vs local scope of trace

For heartbeat info of various adapters, look for Network Interface Monitor (NIMs) file:

`/var/ct/<domain>/log/cthats/nim.cthats.<device>`

Pacemaker

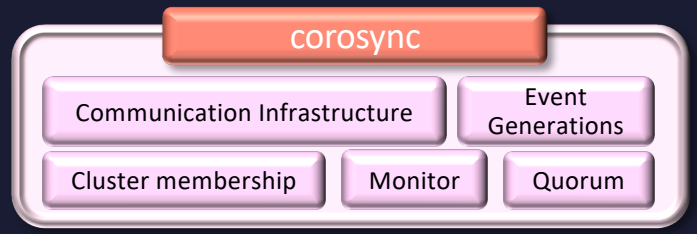


Find domain leader

- `db2cm -list | grep leader`

`/var/log/pacemaker/pacemaker.log, syslog`

Corosync



`/var/log/cluster/corosync.log, syslog`

Performance comparisons between TSA and Pacemaker

1. Reboot: *Pacemaker is faster in recovery from various reboot scenarios*

- In dual reboots scenario with Pacemaker, DB recovery is ~45% faster
- In reboot standby scenario with Pacemaker, Standby reintegration is ~28% faster

2. kill –9 db2sysc: *Pacemaker is faster in recovery even though TSA detects some intermediate events faster*

- In kill –9 Primary scenario, Pacemaker provides ~33% faster DB resource recovery
- In kill –9 Standby scenario, Pacemaker provides ~31% faster instance resource recovery

3. User initiated takeover: Pacemaker is faster overall process

- Provides 24% faster overall process over TSA

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

Agenda

The Foreword

- "Why"

The Preface

- "How"

The Introduction

- Many "Whats"

The evaluation process began in mid 2019 …

Criteria

1. Aim for SINGLE cluster manager for:
   - pureScale, HADR, DPF HA
   - all supported platforms – AIX, Linux
   - On-prem, VM, non-containerized cloud, containerized cloud
2. Minimum dependency on other software owned by others
3. Support Mechanism
4. Rolling update (for pS)
5. Licensing & Cost
6. Reputation & Readiness
7. Ability to integrate with Db2

Candidates

	Intel RHEL & SLES	Linux on Z RHEL & SLES	PPCLE RHEL	AIX	Linux VM	Bare-metal	Container	Rolling Update	Integration with Db2	Cost
Pacemaker	Yes	Yes[1]	Possible[2]	Possible[3]	Yes	Yes	Possible	Yes	Yes	N/A

[1] Db2 built RPMs and validated

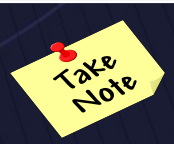
[2] Need to build RPM and validate

[3] Need development & test

What platforms, Db2 releases & deployment env. are supported ?

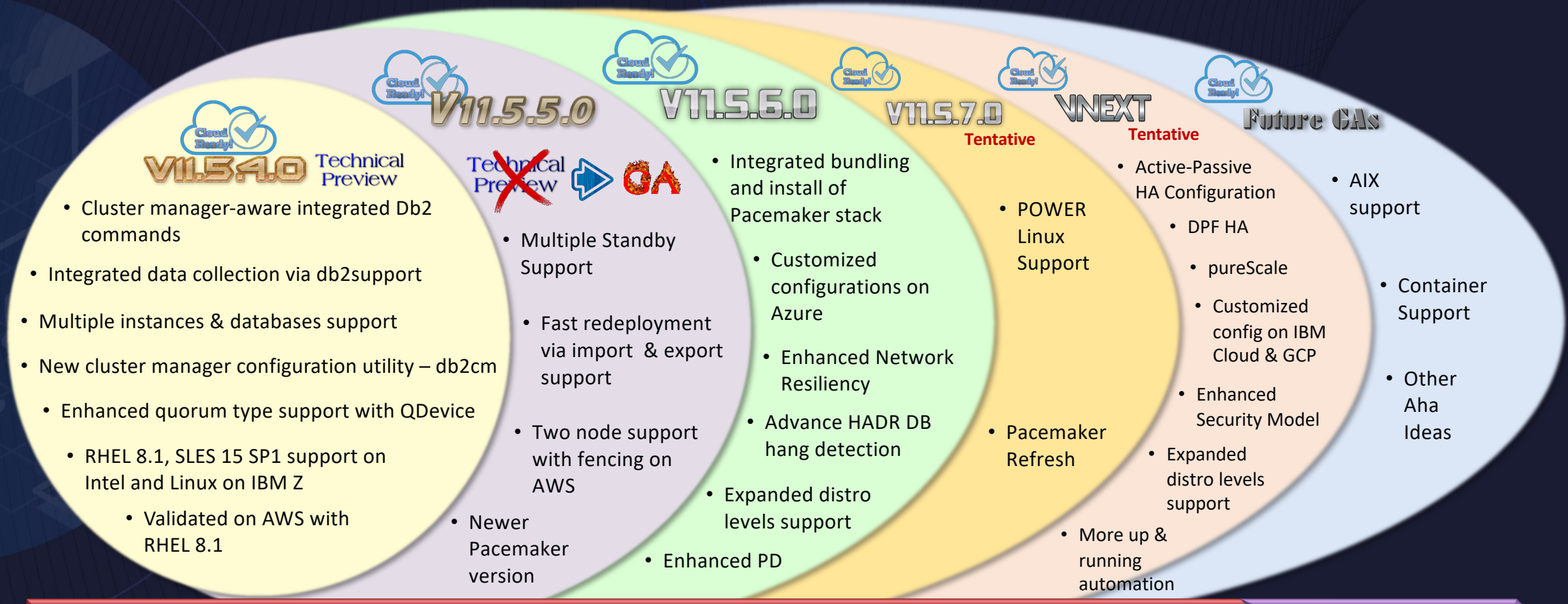
Architecture / Platforms / OS Version	TSA	Pacemaker
Intel / RHEL / 7.x	V11.5.4.0	No Plan
Intel / RHEL / 8.1	V11.5.4.0	V11.5.4.0+
Intel / SLES / 12 SPx	V11.5.4.0	No Plan
Intel / SLES / 15 SPy	No	V11.5.4.0+
Linux on IBM Z / RHEL 8.1	V11.5.4.0	V11.5.4.0+
Linux on IBM Z / SLES 15 SP1	No	V11.5.4.0+
POWER 8 RHEL 7.x	Yes	No Plan
POWER 8 & 9 / RHEL 8.2+	No	In roadmap
POWER 8 & 9 / SLES 15 SPy	No	In roadmap
POWER / AIX / 7.2 TL4	V11.5.4.0	In roadmap

Environments	TSA	Pacemaker
On-premise DC	Yes	Yes
Non-containerized Private Cloud	No	Yes
Non-containerized Public Cloud	No	Yes, validated on AWS and Microsoft Azure on Intel RHEL
Container	No	Not yet.



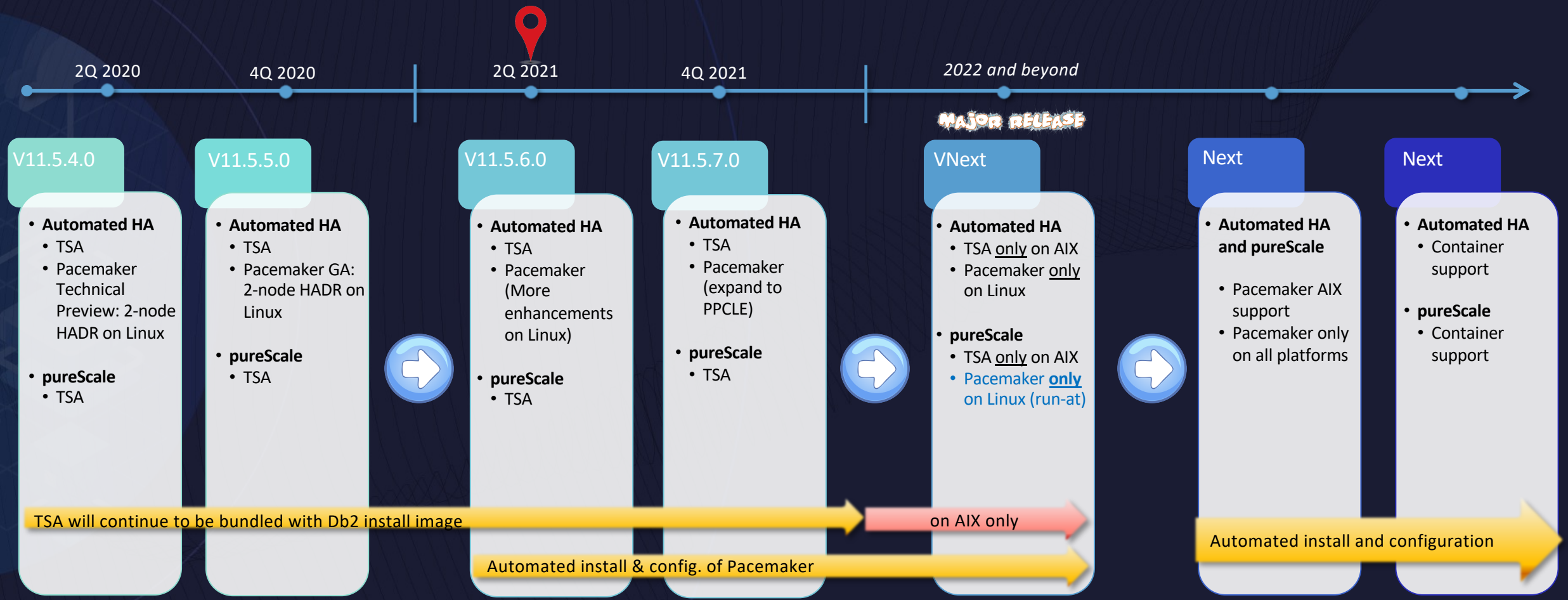
No plan to support Pacemaker as Integrated solution with older version of RHEL (7.x) and SLES (12 SPx)
No plan to support Pacemaker as integrated solution in earlier Db2 releases.

Rapid Agile Delivery Model – Laser-focus on MVP with short dev cycle



Note: Roadmap and content subjected to change

Sneak peek at the roadmap



Note: Roadmap subjected to change

The takeaway …

- All Db2 LUW solutions with TSA as the integrated cluster manager where the corresponding Pacemaker solution is already available are officially in maintenance mode.
- Future enhancements will be considered in the Pacemaker path as a priority.
- Steps for converting from TSA to Pacemaker will be provided as new solutions with Pacemaker are rolled out.
  - The conversion step for 2-node configuration is already available (see later slide)
  - The same will be provided in subsequent releases when DPF HA and pureScale support with Pacemaker are made available.

Agenda

The Foreword

- "Why"

The Preface

- "How"

The Introduction

- Many "Whats"

Packaging of Cluster Software

V11.5.4.0 (Technical Preview) + V11.5.5.0 (GA):

- Not bundled with Db2 install image (yet)
- The Pacemaker and all dependent cluster software will be packaged in a tar file, available for download in an IBM hosted website (more detail later)
- Separate install
  - provides an easy way to provide fixes and release new support without engine changes (no need for special builds)
- Automated setup and configuration
  - Through new Db2 cluster manager utility – db2cm (more detail later)

V11.5.6.0:

- Full bundling and install integration – Pacemaker stack is part of the Db2 install image.
- Download site is used to host cloud specific fencing agents

Integrated Pacemaker install (V11.5.6.0)

- Pacemaker Software Stack available as part of base Db2 install image!
  - MRS site will only hold cloud specific RPMs (may change in future)

- Fresh install or update: single command installs Pacemaker by default:
  - `db2_install -y -b -/opt/ibm/db2/V11.5 -p SERVER`
  - `installFixPack -y -b /opt/ibm/db2/V11.5 -p /opt/ibm/db2/V11.5.6`

- Both db2_install & installFixPack handle Pacemaker version upgrade automatically:
  - Upgrade when current Pacemaker is Db2 supplied and has lower version than the target one
  - Skip if current Pacemaker:
    - is not Db2-supplied
    - is Db2-supplied but is already at higher version

```
Task #33 start
Description: TSA
Estimated time 300 second(s)
Task #33 end

Task #34
start Description: Pacemaker
Estimated time 300 second(s)
Task #34 end
.
.
.
The execution completed Successfully
```

```
WARNING: DBI1986E There is already a Pacemaker cluster
manager installed on the system that is not provided by
IBM. Remove the current installation of Pacemaker before
proceeding with your IBM-provided Pacemaker installation.
```

Integrated Pacemaker install (V11.5.6.0) – cont'd

- Skip installation of Pacemaker during initial install with –NOPCMK flag
  - db2_install -p server -b /opt/ibm/db2/V11.5 -NOPCMK
- If –NOPCMK flag was used to install Db2, Pacemaker can be installed later using db2installPCMK script included in the install image.
  - <Db2_install_image>/universal/db2/<platform>/pcmk/db2installPCMK –i
- No longer require EPEL or Backport repositories to be setup for Pacemaker installation.
  - Simpler setup, less maintenance going forward.

```
dnf install https://dl.fedoraproject.org/pub/epel/epel-release-latest-8.noarch.rpm
zypper addrepo -fc http://download.opensuse.org/repositories/openSUSE:/Backports:/SLE-15-SP1/standard/openSUSE:Backports:SLE-15-SP1.repo
```

Note: *Fencing agents will be available separately from the MRS site.*

Db2 Pacemaker Documentation

Pacemaker (Linux)

- **Database administration**
  - + Db2 data servers
  - + Instances
  - + Databases
  - + Database objects
  - + Data movement utilities and reference
  - **High availability**
    - + Outages
    - **High availability strategies**
      - Redundancy
      - Failover
      - **Clustering**
        - **Supported cluster management software**
          - **Pacemaker (Linux)**
          - IBM PowerHA SystemMirror for AIX
          - IBM Tivoli System Automation for Multiplatforms (Linux and AIX)
          - Microsoft Failover Clustering support (Windows)
          - VERITAS Cluster Server
      - + Database logging

Troubleshooting Pacemaker

- Troubleshooting Pacemaker
  - Db2 instance fails to restart automatically after a failure
  - Db2 HADR database pair both assume primary role
  - Database resource shows both HADR databases as standby
  - Database resource stuck in stopped state

NEW V11.5.5.0

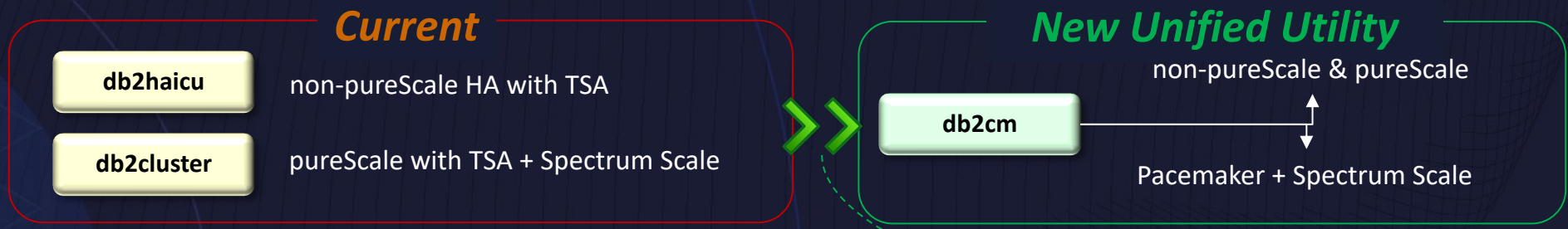
Installing the Pacemaker cluster software stack

- **Database fundamentals**
  - **Installing**
    - Requirements for Db2 products
    - + Installing Db2 database servers
    - + Installing IBM Data Server drivers and clients
    - + Installing a Db2 pureScale environment
    - + Installing the Db2 Partitioned Database Environment
    - + Installing Db2 products and features using a response file
    - **Installing the integrated cluster manager**
      - + Installing and upgrading SA MP with the Db2 installer
      - **Installing the Pacemaker cluster software stack**
    - + Installing IBM Data Studio
    - Using the Guardium Installation Manager Client

Integrated solution using Pacemaker

- **Integrated solution using Pacemaker**
  - Pacemaker base component
    - Networks in a Pacemaker cluster
    - Quorum devices support on Pacemaker
  - Prerequisites for an integrated solution using Pacemaker
  - db2cm - Db2 cluster manager utility
  - Configuring a clustered environment using the db2cm utility
    - Installing the Pacemaker cluster software stack
    - Install and configure a QDevice quorum
    - Public cloud vendors supported with Db2 Pacemaker
      - Setting up two-node HADR with fencing on Amazon Web Services (AWS)
    - Removing a cluster domain
    - Maintaining a Pacemaker cluster domain
      - User initiated takeover
      - User initiated takeover by force
      - Add a HADR database resource to the resource model
      - Delete an existing HADR database resource from the resource model
      - Associate a primary VIP with an existing HADR database of an instance
      - Disassociate a primary VIP with an existing HADR database of an instance
      - Associate a standby VIP with an existing HADR database of an instance for read-on-standby
      - Disassociate a standby VIP with an existing HADR database of an instance
      - Remove all resources related to the public Ethernet adapter device on a host in the resource model
      - Remove all resources related to an instance in the resource model
      - Remove an automated HADR cluster with Pacemaker
      - Backup cluster configuration information
      - Restore from a saved Pacemaker cluster configuration
    - + Troubleshooting Pacemaker
    - Restrictions on Pacemaker
    - Converting an existing Tivoli SA MP cluster to a Pacemaker cluster
    - Converting an existing Pacemaker cluster to a Tivoli SA MP cluster

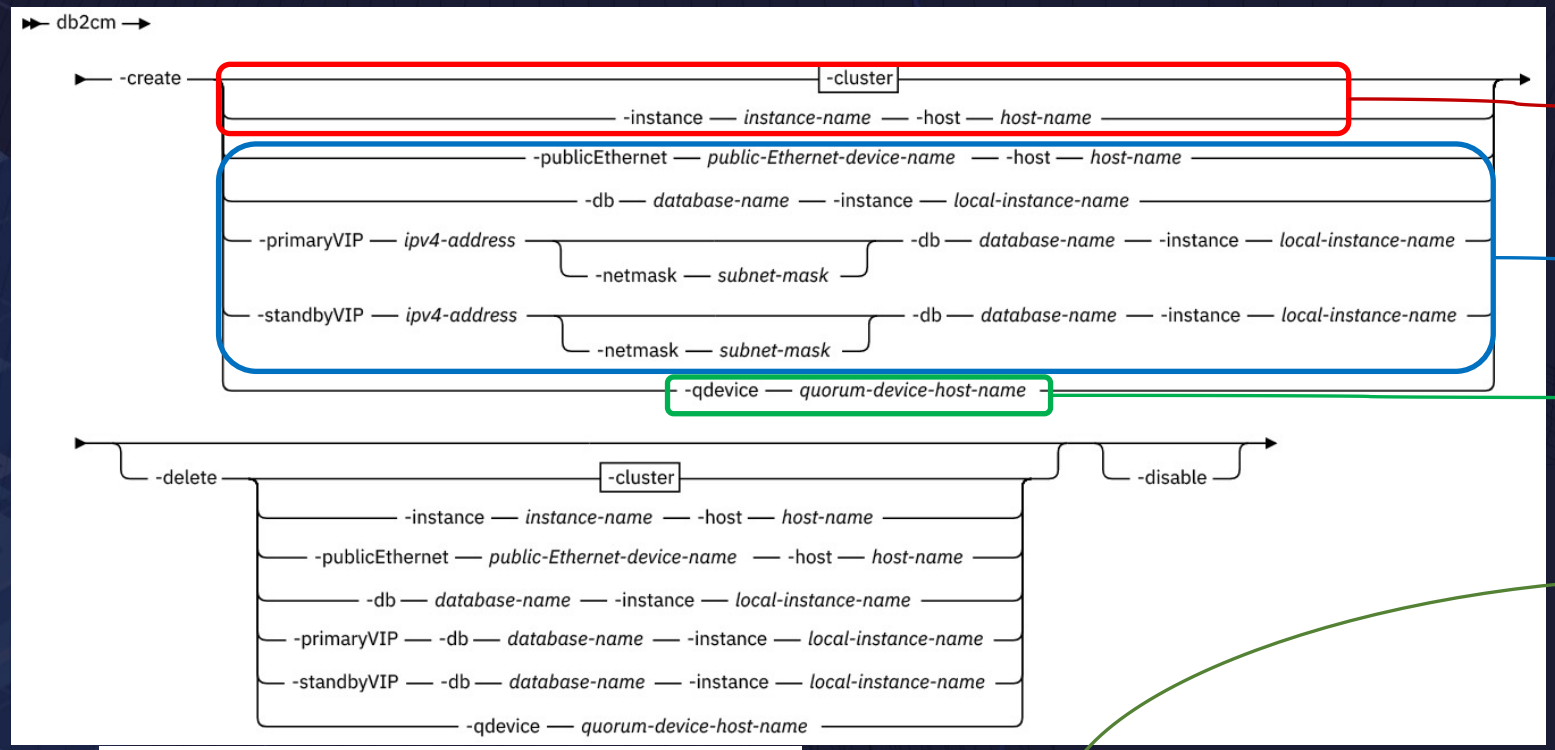
New integrated configuration tool – *db2cm*



Stability	• **ZERO** chance to regress db2haicu and db2cluster
Usability	• more "db2cluster" like. Option-based, not menu- driven utility
Clear distinction between the "OLD" Vs "New"	• Easier to deprecate (and remove) in the future

- Convergence is the long-term strategy.
- In V11.5.4.0 and V11.5.5.0, db2cm is only used for the integrated Pacemaker solution.

db2cm - syntax diagram

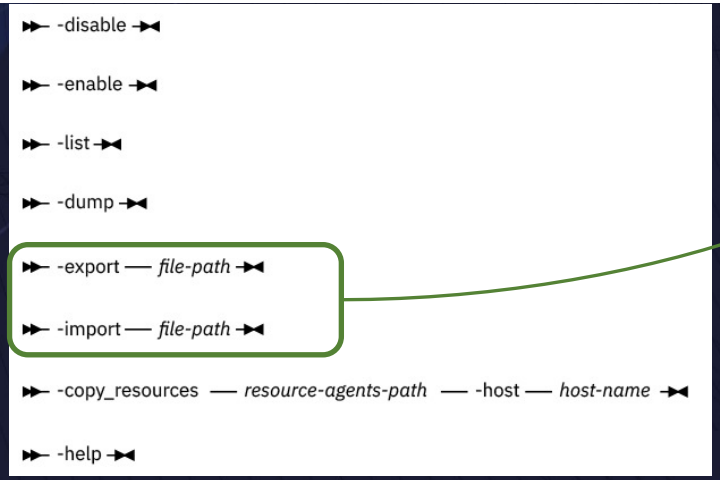


Domain

Instance, HADR DB, Ethernet Resources

Quorum

Export / Import cluster configuration



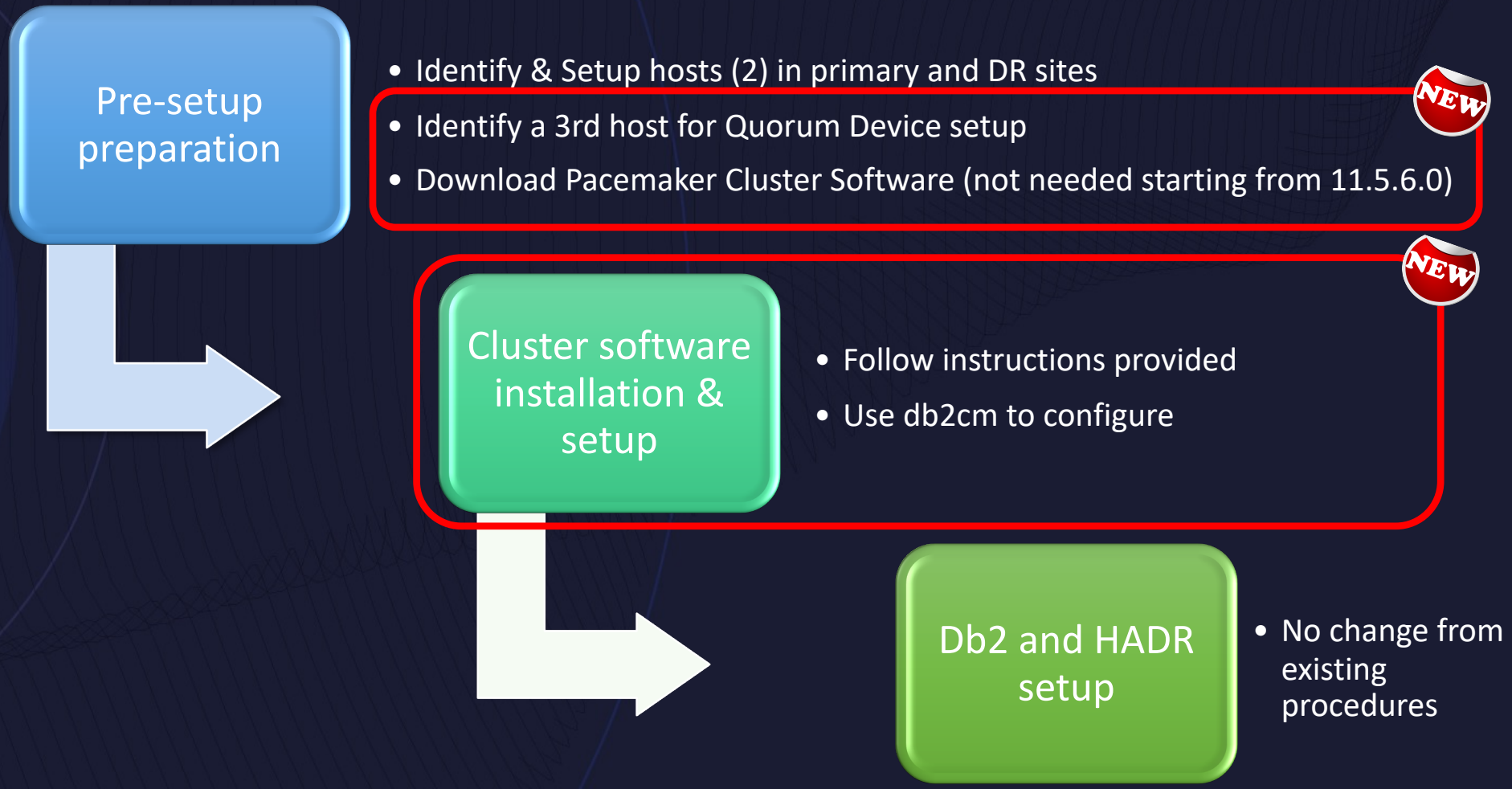
Take Note

db2cm is part of Db2 integrated solution with Pacemaker.

Every db2cm invocation checks for Db2 version of Pacemaker and fails if it isn't. (recall db2cm should ONLY be used on integrated solution)

```
> db2cm -create -cluster -instance …

Please use Pacemaker and Corosync packages provided by IBM.
These can be found at: https://www-
01.ibm.com/marketing/iwm/platform/mrs/assets?source=mrs-
db2pcmk"
```

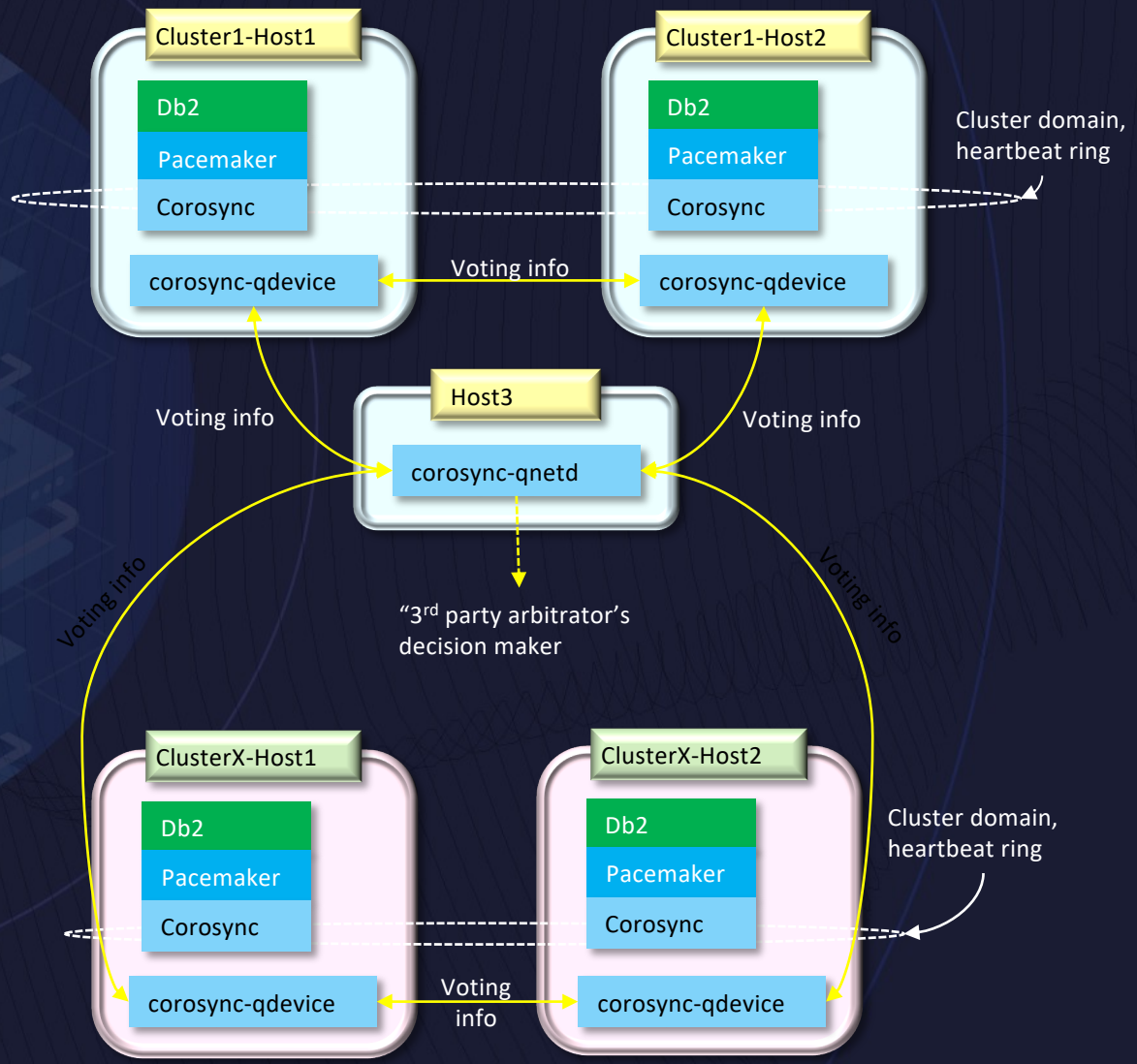
Usage Scenario: General flow of new installation & setup



Quorum Support

- No IP Tiebreaker and disk tiebreaker support in Pacemaker
- Pacemaker recommends Qdevice for reliable Quorum
  - Requires dummy node to run arbitrator daemon
    - No need to install Db2 or setup Pacemaker in the dummy node. (more details next slide)
  - Azure also recommends a 3rd host to provide STONITH Block Device (SBD) fencing.
- Qdevice quorum is the recommendation for 2-node HADR setup

Qdevice Quorum & Qnetd Overview



Qdevice Arbitrator Node Details

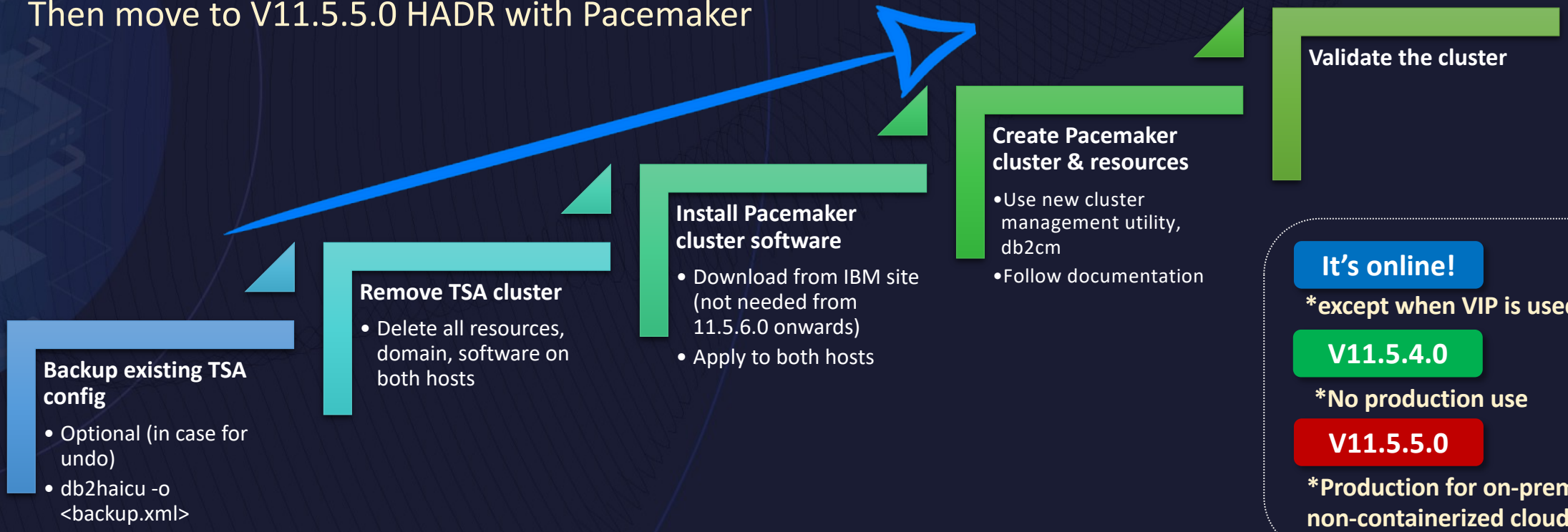
- Flexible in platform, architecture
  - e.g. Validated using the same host for RHEL, SLES clusters on Intel and Z.
- Small memory, disk footprint
  - Only need to install corosync-qnetd RPM
  - No need to install Db2 or Pacemaker
  - Not part of the Pacemaker cluster
- TCP/IP accessible from all hosts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

Usage Scenario: Convert from TSA to Pacemaker

Typical "in-release" conversion scenario, start with one of the following:

- Mod pack upgrade from V11.5.0.0 HADR with TSA to V11.5.4.0 HADR with TSA
- Mod pack upgrade from V11.5.4.0 HADR with TSA to V11.5.5.0 HADR with TSA
- New V11.5.4.0/V11.5.5.0 HADR instance w/ TSA

Then move to V11.5.5.0 HADR with Pacemaker



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**

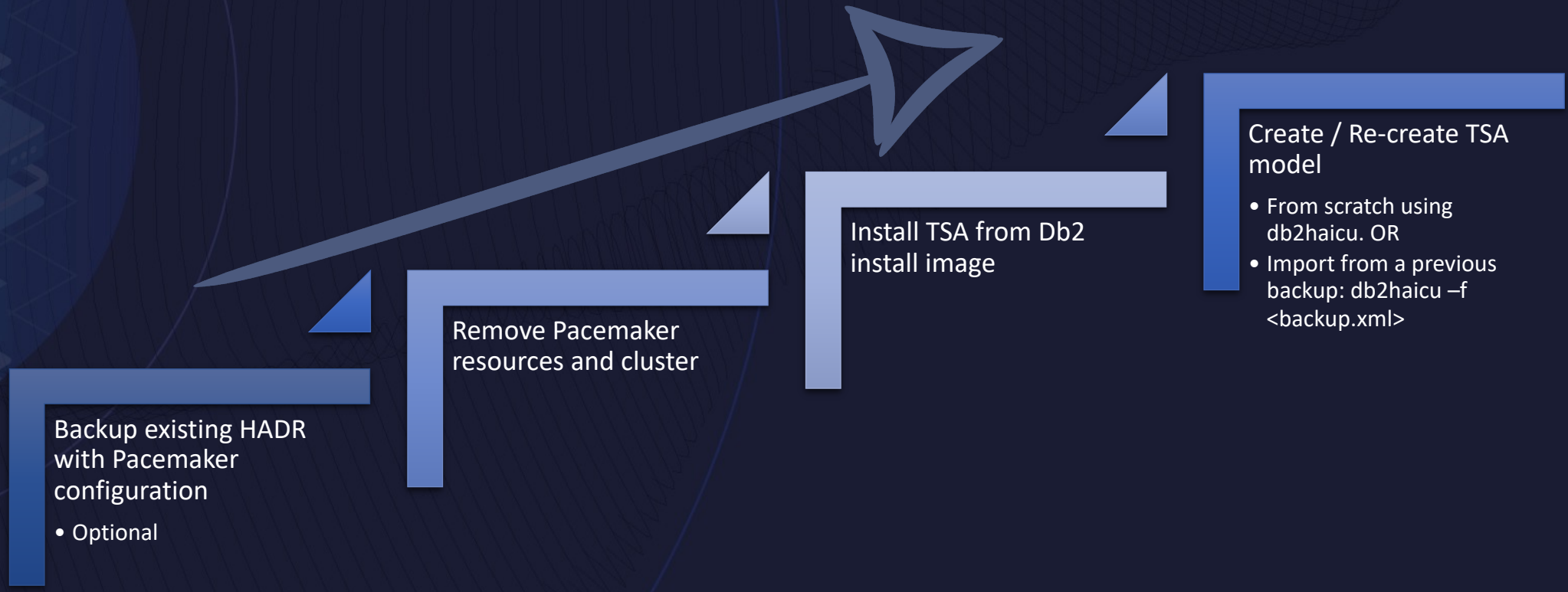
Usage Scenario: Convert from Pacemaker back to TSA

Assume “in-release” conversion scenario:

- From V11.5.4.0 HADR with Pacemaker to V11.5.4.0 HADR with TSA
- From V11.5.5.0 HADR with Pacemaker to V11.5.5.0 HADR with TSA

It’s online!

***except when VIP is used**



Usage Scenario: Fast re-deployment on same hardware

- **Backup configuration**

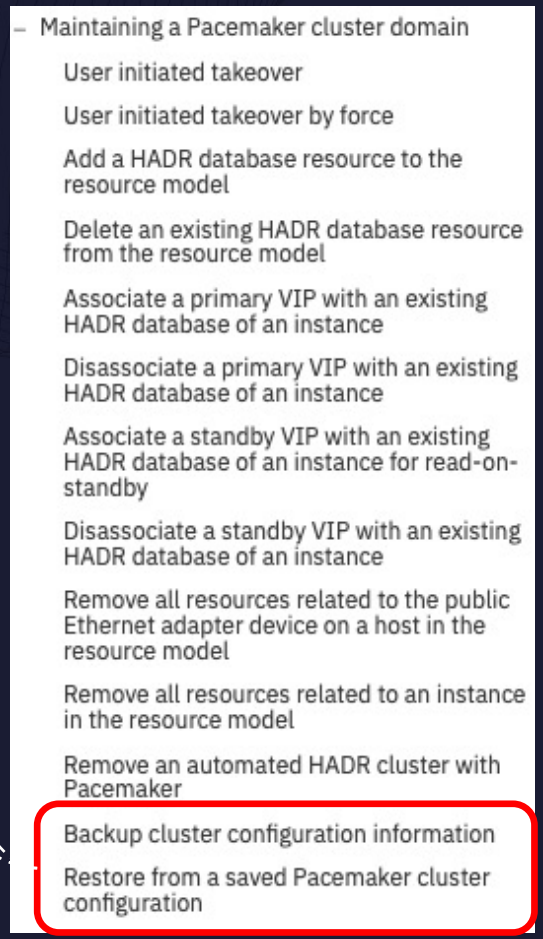
```
[root@jesting1]$ /home/db2inst1/sqllib/adm/db2cm -export /tmp/backup.conf
Exporting configuration to /tmp/backup.conf

[root@jesting1]$ ls -la /tmp/backup.conf
-rw-r--r-- 1 root root 12888 Sep 1 14:22 /tmp/backup.conf
```

- **Restore configuration** (need to clean up existing environment via `db2cm -delete -cluster` first)

```
[root@jesting1]$ /home/db2inst1/sqllib/adm/db2cm -import /tmp/backup.conf
Importing configuration from /tmp/backup.conf
Cluster created successfully.
```

Fast deployment on NEW hardware is possible:

- Requires manual changes to exported file
- Example available in technote off KC

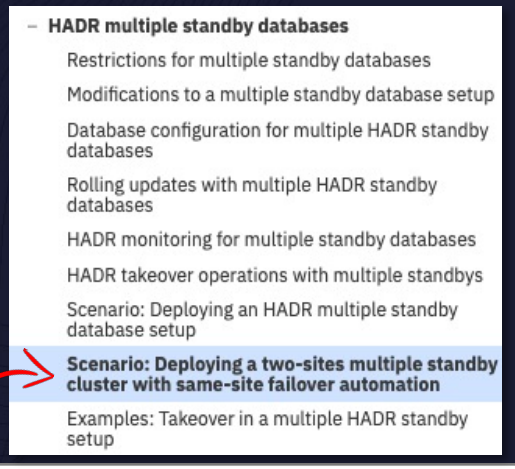


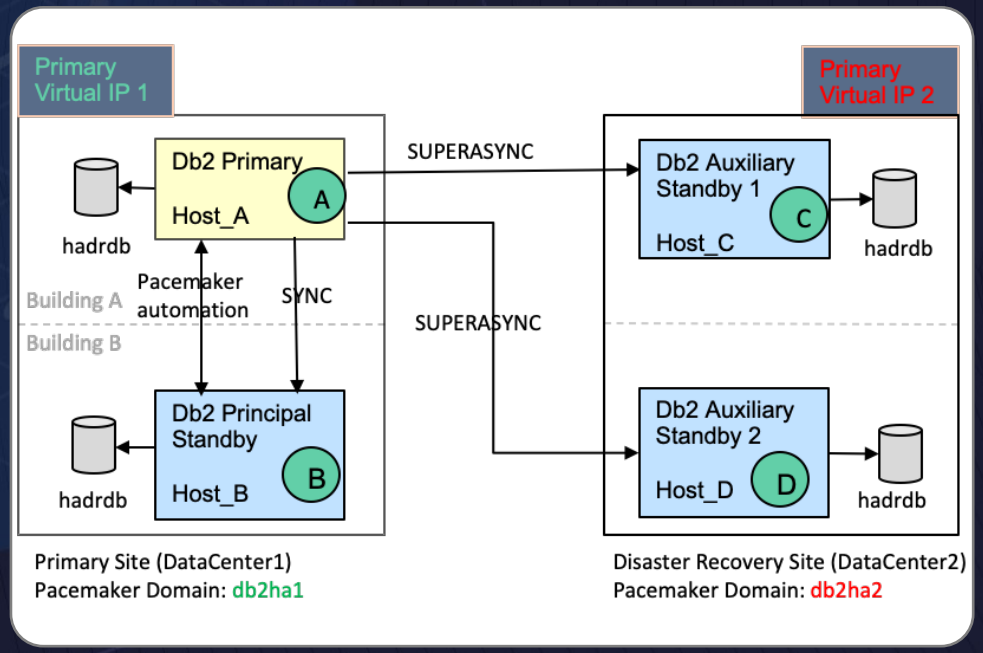
Link to KC

Multiple Standby Support

- Same support as with TSA solution
  - Up to 3 standbys for each HADR DB
  - Auxiliary standbys can be in 1 or 2 sites that is same or different from primary
  - Automatic failover supported between Principal Primary and Principal Standby

Step-by-step setup procedures on each host added to KC

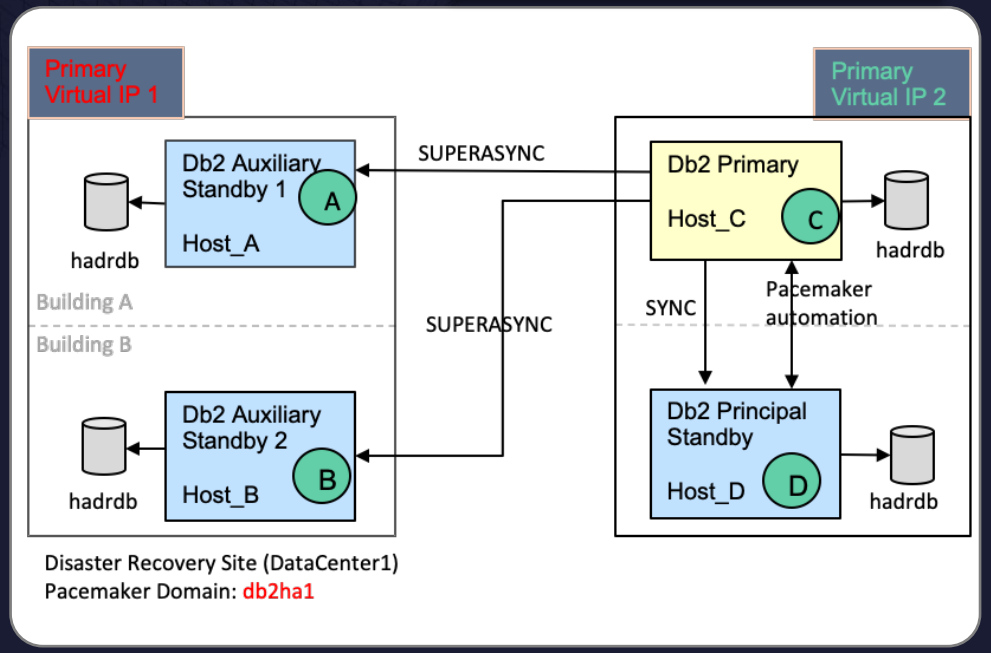




On host C or D, run `TAKEOVER HADR DB`

Pick the host with the most up to date log files

```
db2pd -hadr -db hadrdb | grep STANDBY_LOG_FILE,PAGE,POS
```



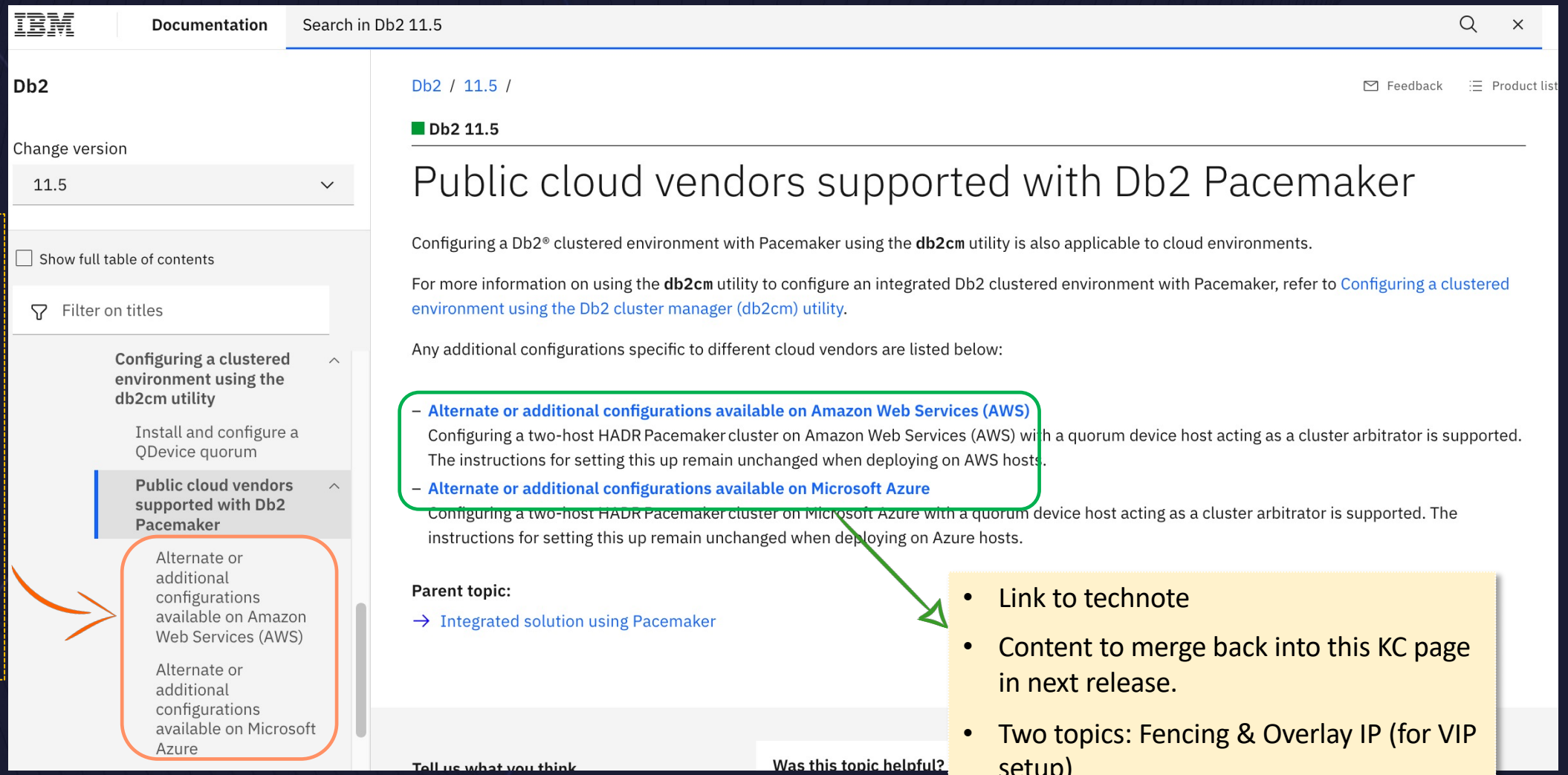
Best Practice Configuration

Specific or alternate configurations on public cloud

Link to KC

New subsection

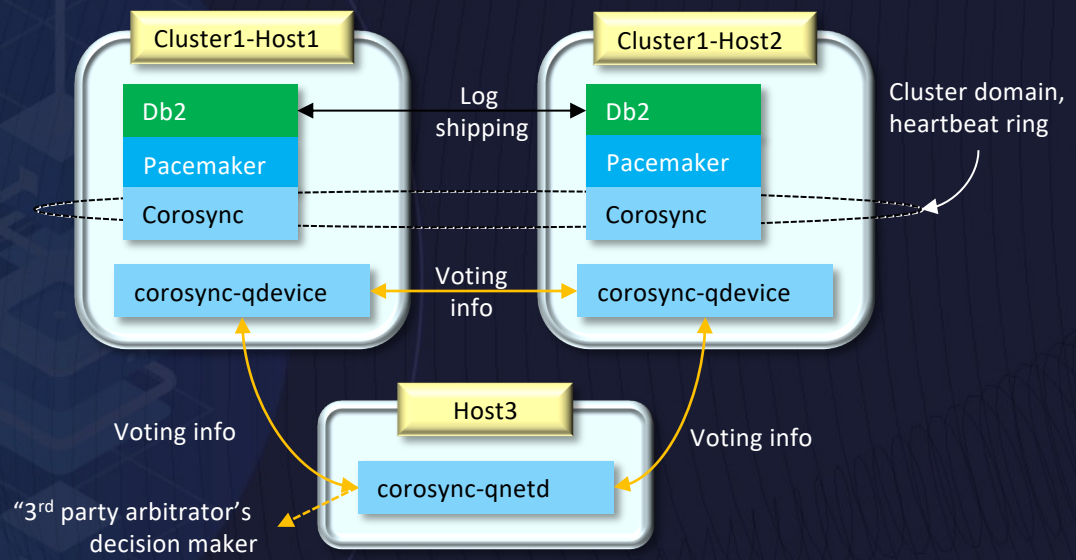
- AWS in V11.5.5.0
- Azure in V11.5.6.0
- Will grow to include IBM Cloud, GCP, etc. next release
- Content expected to include only the “delta”. Not duplicate the end-to-end instructions as for on-prem



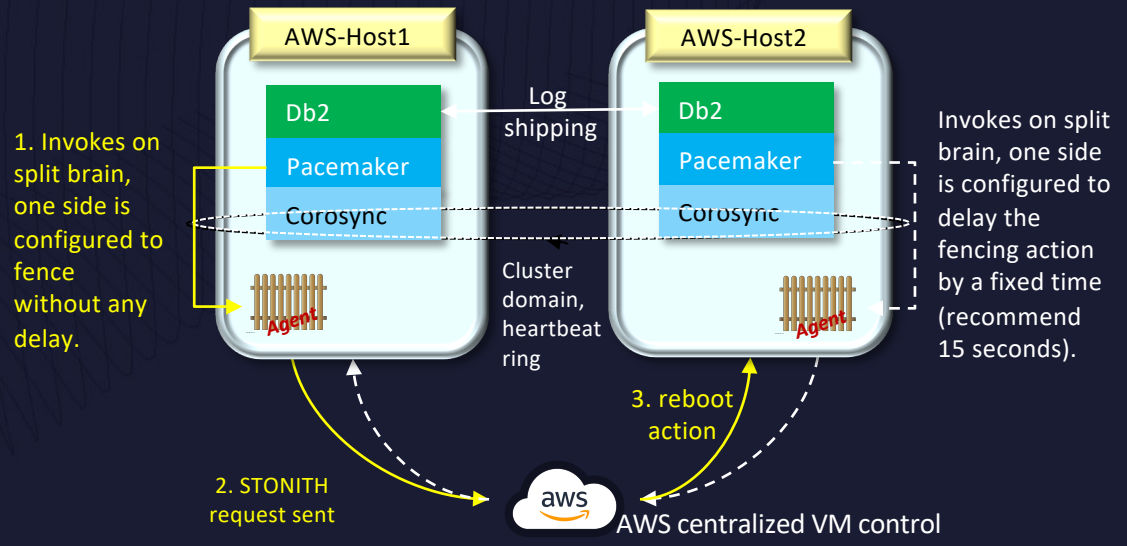
- Link to technote
- Content to merge back into this KC page in next release.
- Two topics: Fencing & Overlay IP (for VIP setup)

New 2-node only support with fencing on AWS

Best practice "2+1" node configuration
(introduced in technical preview)



True 2-node configuration with fencing
(validated on AWS with fence_aws agent)



Pros

- No need of 3rd VM (Lower on-going cost in public cloud)

Cons

- Longer recovery time (internal test showed up to 6 times slower than with quorum device)
- Additional configuration - longer HADR_PEER_WINDOW for each DB (recommend at least 300 seconds)

Summary

- **Key topics covered in today's session:**
  - Reasons why Pacemaker is a superior solution
  - Overview of Pacemaker Architecture
  - Quorum mechanism
  - Deployment, configurations
  - Multiple Standby, Fast redeployment, Cloud config.
  - Sneak peek at the roadmap

Speaker: Hao Qi
Company: IBM
Email Address: haoqi@ca.ibm.com