Pacemakerで学ぶ HAクラスタ

OSC2019 Tokyo/Spring
2019/2/23

Linux-HA Japan
松浦　健太

はじめに：本日の話の流れ

1. 可用性とクラスタ
   - 可用性とは、可用性を向上させるためには
   - 「クラスタ」の必要性と分類について

2. Pacemakerとは
   - Pacemakerの概要と動作

3. いろんなHAクラスタ
   - HAクラスタでのサーバ構成とデータ管理

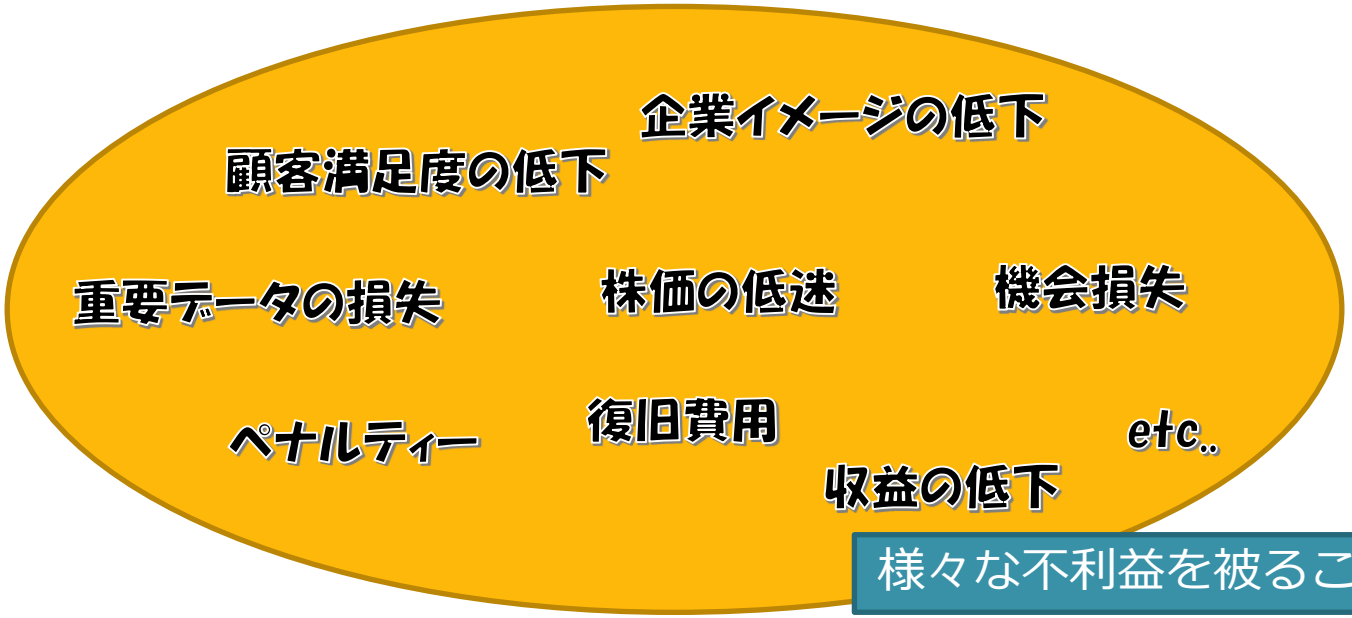
4. まとめ

可用性とクラスタ

可用性とは

可用性(Availability)＝サービス継続性

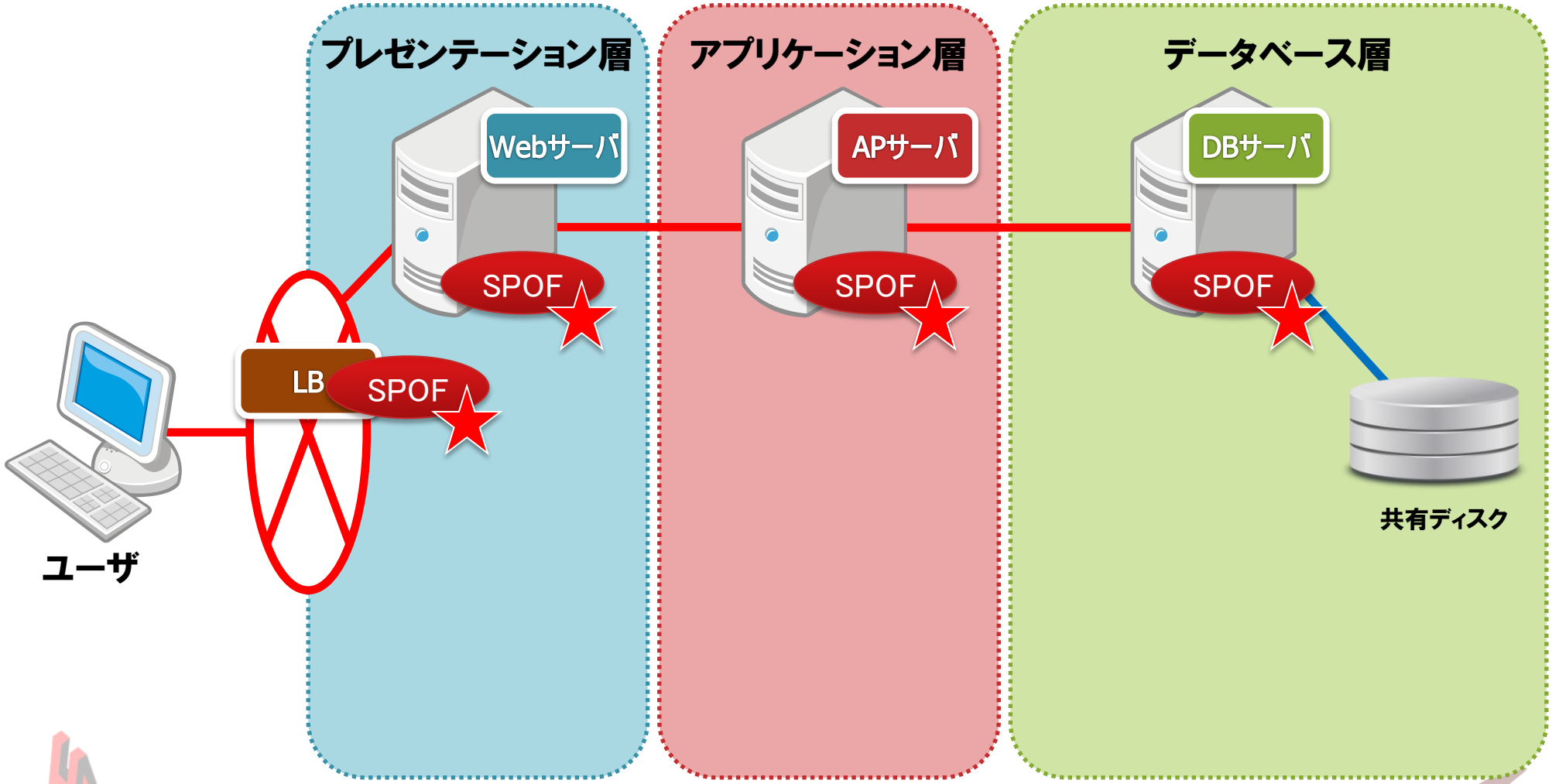
もしサービスが止まってしまったら？？



→ 可用性を高め「止まらないサービス」の実現を目指す

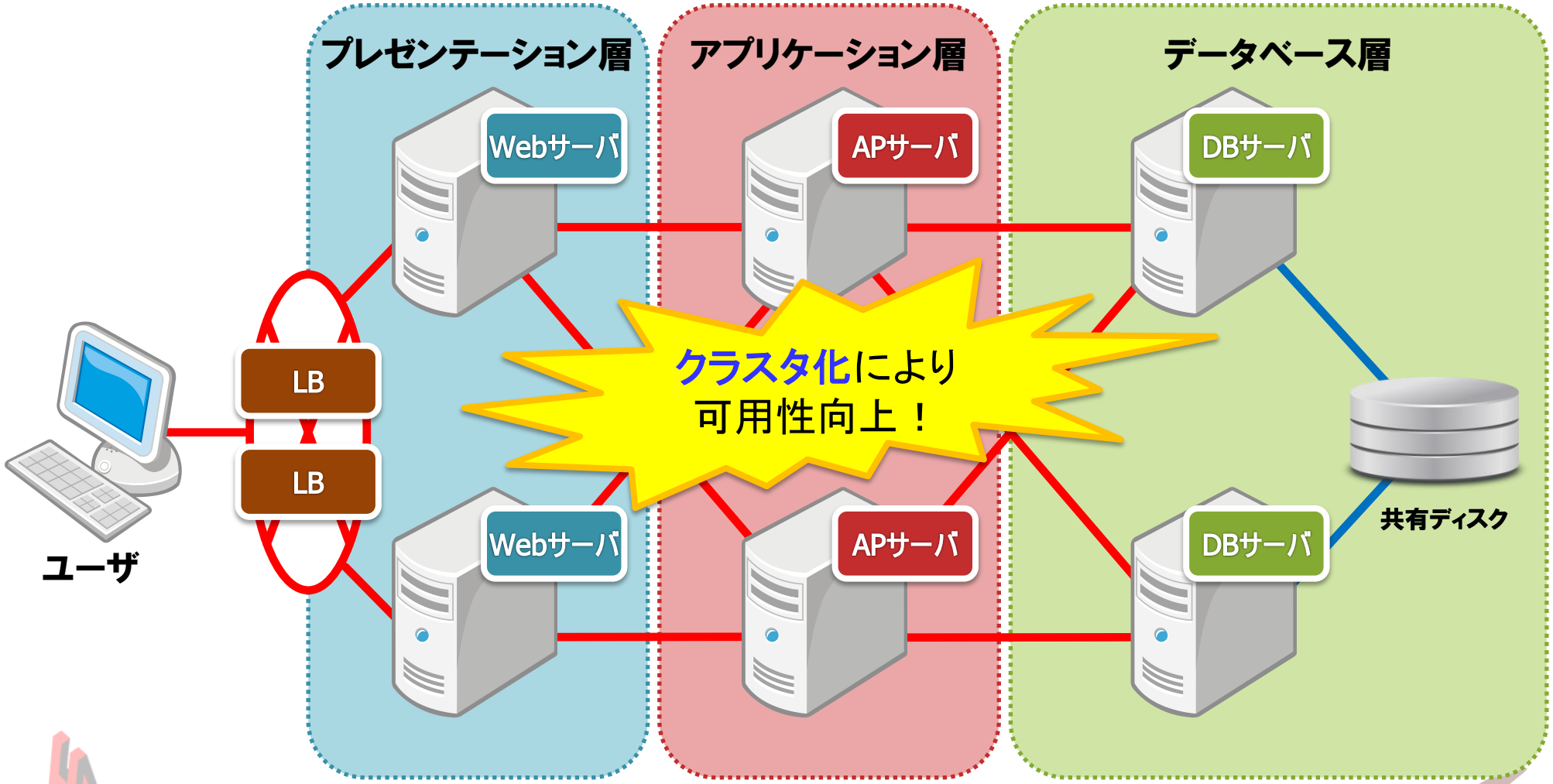
可用性の向上

可用性を高めるためには単一障害点(SPOF : Single Point of Failure ある一点が故障した場合にサービスが停止してしまう部位)の除去が重要



可用性の向上

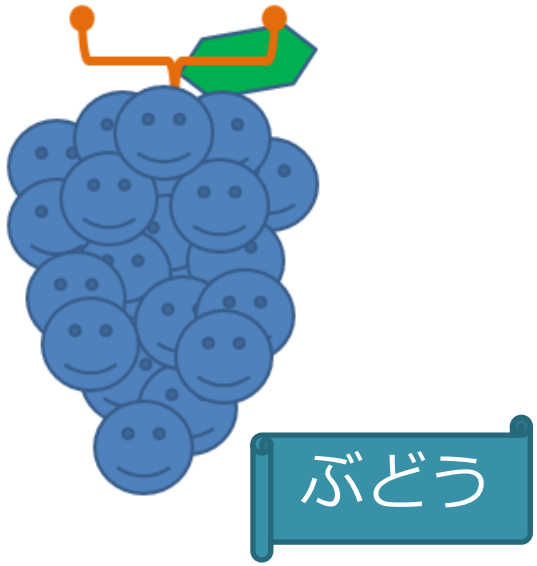
可用性を高めるためには単一障害点(SPOF：Single Point of Failure ある一点が故障した場合にサービスが停止してしまう部位)の除去が重要

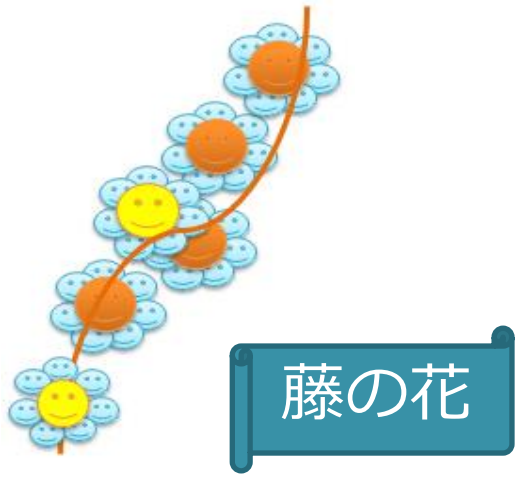


クラスタとは

ところで、クラスタとは？

- 語源は果実や花の房
- 同じようにまとまっているもののことを指す





- IT業界でいうクラスタとは、複数のコンピュータを連携させ、
  全体で一つのコンピュータのように振る舞わせる仕組みを指す

クラスタの分類

負荷分散クラスタ

同一のサービスを提供する複数台のサーバに対して処理を分配する
そのため、一台が停止しても残りのサーバでサービスを継続することができ、
かつ一台のサーバでは得られなかった処理性能を確保することが可能

HA(High Availability：高可用)クラスタ

メインでサービスを提供するサーバ(稼働系,現用系,Activeなどと呼ばれる)と、
それをバックアップするサーバ(待機系,予備系,Standbyなどと呼ばれる)で
構成され、現用系に障害が発生した際に予備系でサービスを引き継ぐことで、
サービスの停止時間を短くする

可用性を向上できるのはこの2つ

HPC(High Performance Computing)クラスタ

複数台のコンピュータを結合させて演算処理を分担し、
一台のコンピュータでは得られない高性能を確保することを目的とする

クラスタの分類

負荷分散クラスタ

同一のサービスを提供する複数台のサーバに対して処理を分配する
そのため、一台が停止しても残りのサーバでサービスを継続することができ、
かつ一台のサーバでは得られなかった処理性能を確保することが可能

HA(High Availability：高可用)クラスタ

メインでサービスを提供するサーバ(稼働系,現用系,Activeなどと呼ばれる)と、
それをバックアップするサーバ(待機系,予備系,Standbyなどと呼ばれる)で
構成され、現用系に障害が発生した際に予備系でサービスを引き継ぐことで、
サービスの停止時間を短くする

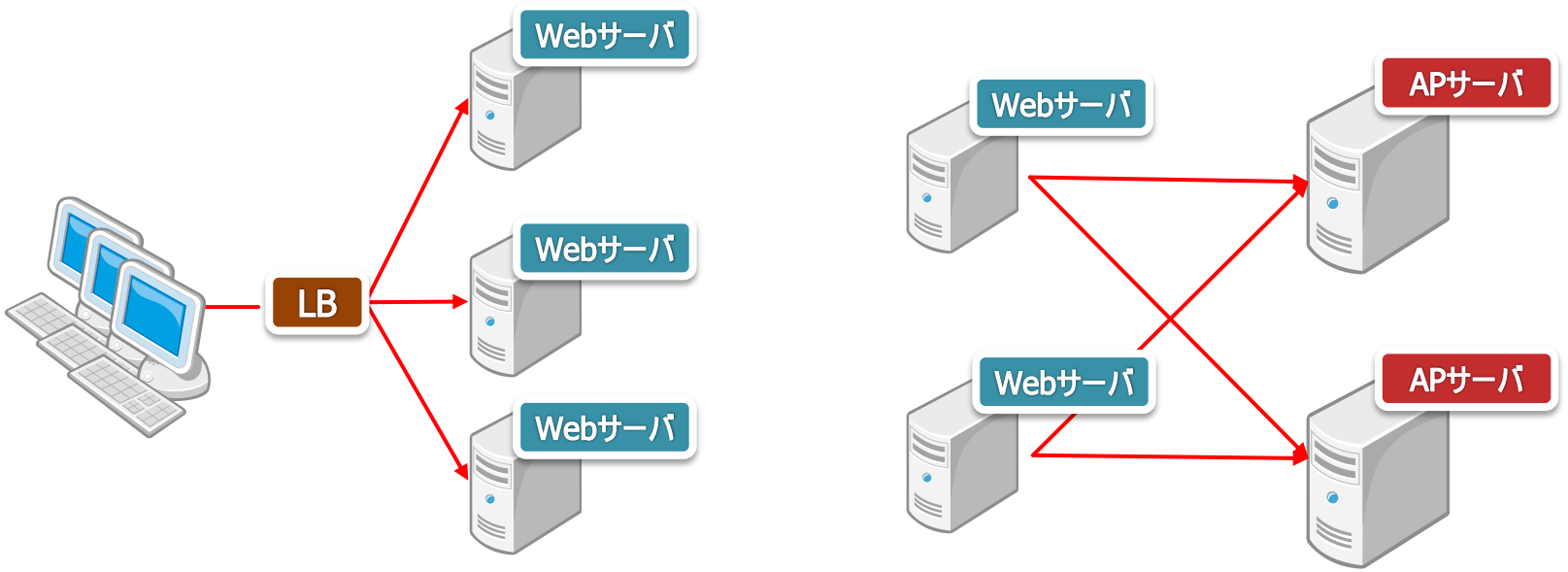
HPC(High Performance Computing)クラスタ

複数台のコンピュータを結合させて演算処理を分担し、
一台のコンピュータでは得られない高性能を確保することを目的とする

負荷分散クラスタ概要

□ 基本構成

- 同じサービスを提供する複数のサーバ群に対して処理を分散させ、1台あたりのサーバ負荷を低減させる
- LB(ロードバランサ)によるWebサーバへの負荷分散や、WebサーバによるAPサーバへの負荷分散処理等、複数のパターンがある



クラスタの分類

◘ 負荷分散クラスタ

同一のサービスを提供する複数台のサーバに対して処理を分配する
そのため、一台が停止しても残りのサーバでサービスを継続することができ、
かつ一台のサーバでは得られなかった処理性能を確保することが可能

◘ HA(High Availability：高可用)クラスタ

メインでサービスを提供するサーバ(稼働系,現用系,Activeなどと呼ばれる)と、
それをバックアップするサーバ(待機系,予備系,Standbyなどと呼ばれる)で
構成され、現用系に障害が発生した際に予備系でサービスを引き継ぐことで、
サービスの停止時間を短くする

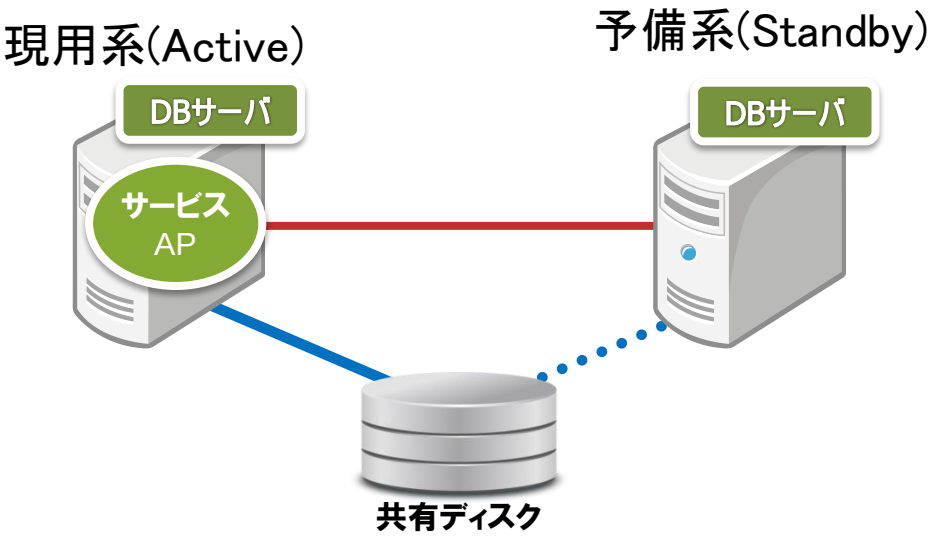
◘ HPC(High Performance Computing)クラスタ

複数台のコンピュータを結合させて演算処理を分担し、
一台のコンピュータでは得られない高性能を確保することを目的とする

HAクラスタ概要

□ 基本構成

- サービス中のサーバに故障が発生した場合、他のサーバに処理を引き継ぐ
- サービスを提供する現用系サーバと待機状態にしておく予備系サーバから成るActive-Standby構成(Act-Sbyと書くことも)が一般的
- 同等の性能を有するサーバを2台配置し、サーバ間でデータの共有が必要な場合は共有ディスク※を使用

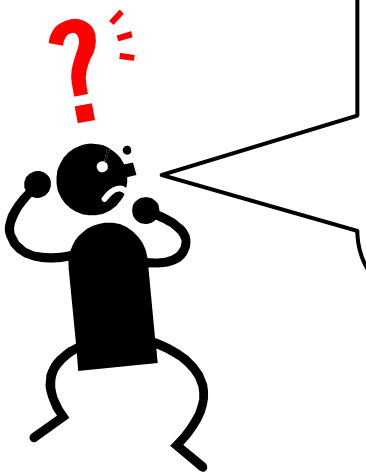


※ 共有ディスクを用いる代わりに、ソフトウェアの機能でデータを他のサーバへレプリケーションすることでデータを共有する構成もある

負荷分散クラスタとHAクラスタの違い

負荷分散クラスタは負荷の分散をしつつ可用性を高める。

一方、HAクラスタは可用性を高めるだけ・・・。

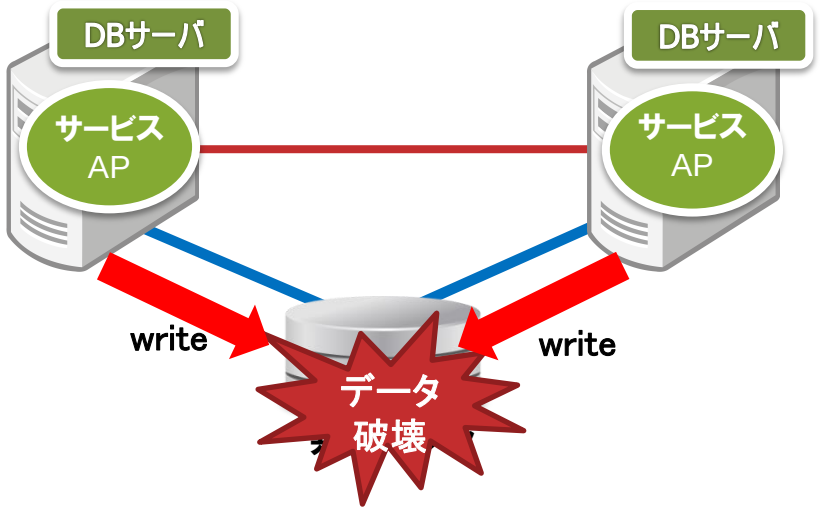


もう負荷分散クラスタで
良くない？

負荷分散クラスタとHAクラスタの違い

負荷分散クラスタとHAクラスタはサービスを2重起動するか否かで
使い分ける。

DBを2重起動した場合



HAクラスタなら2重起動をしないため問題ない

次章からは
このHAクラスタについて
「Pacemaker」を例に
説明します！

Pacemakerとは

Pacemakerとは

Pacemaker(ペーすめーかー)とはオープンソースで
開発されている高可用(High Availability)クラスタソフト

- Heartbeatの後継として開発されたソフトウェアであり、ソースはGitHubで管理
  - ClusterLabs：https://github.com/ClusterLabs
- バイナリファイルは以下より入手可能
  - Linux-HA Japanプロジェクト：http://linux-ha.osdn.jp/wp/

HAクラスタソフト

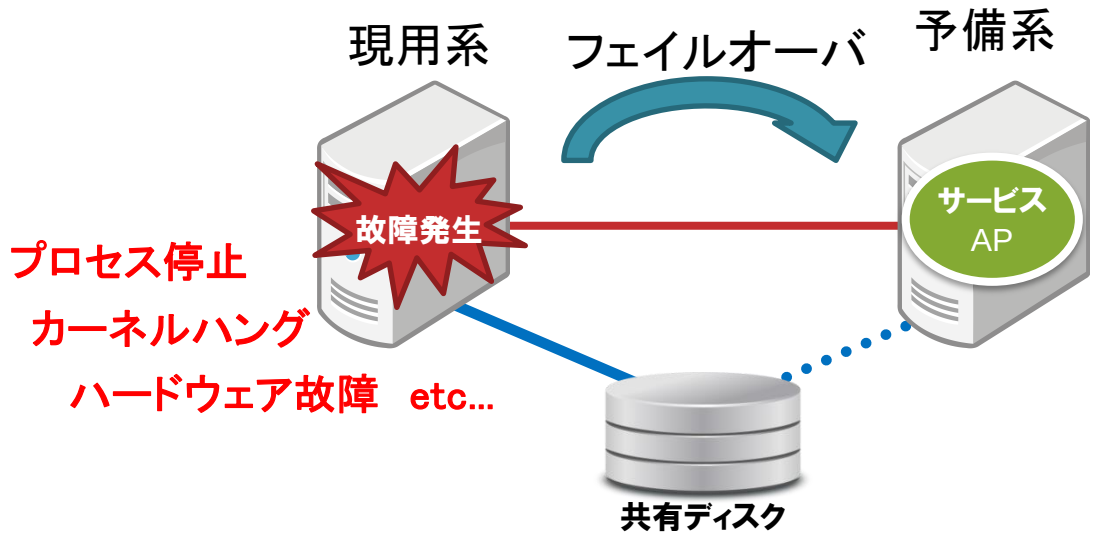
HAクラスタの管理を自動化するソフトウェアを**HAクラスタソフト**という。

□ HAクラスタソフトの主な機能

✓ 故障検知とサービス引継ぎ

・常時ハードウェア及びサービスアプリケーションの稼動状態を監視

・現用系サーバ上で故障を検知した場合、
予備系でアプリケーションを起動してサービスを継続→「フェイルオーバ」という

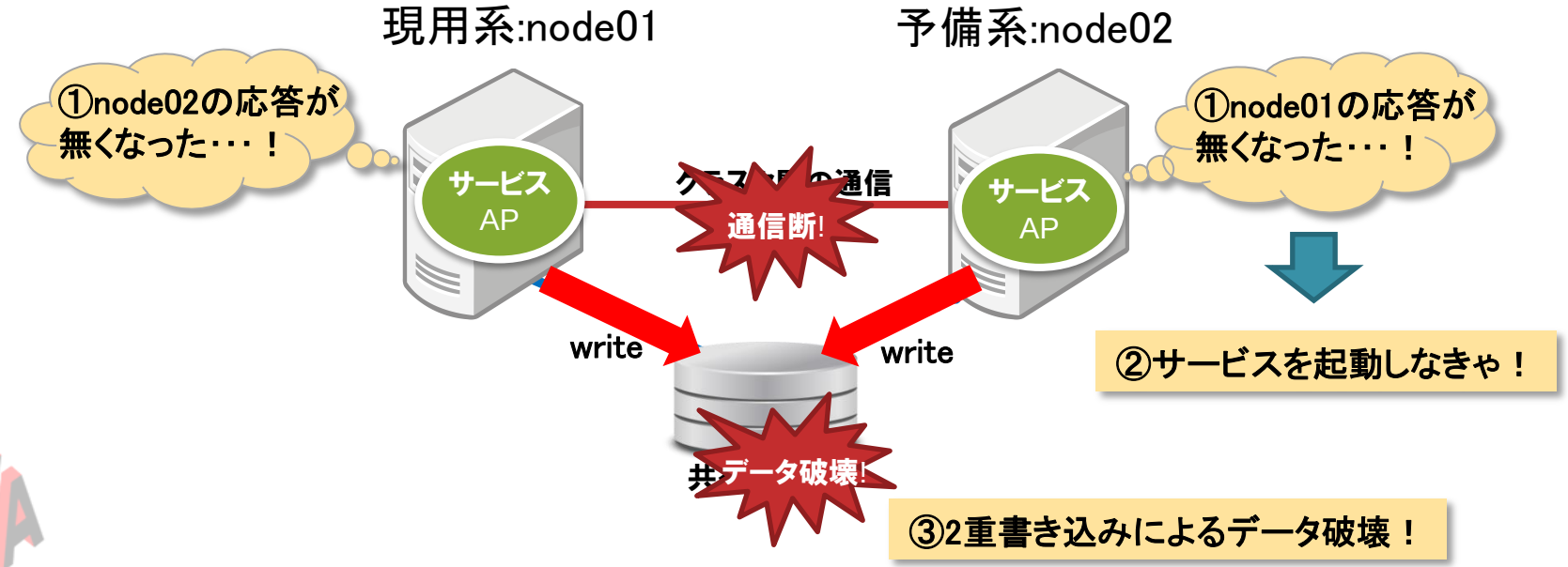


HAクラスタソフト

✓ 排他制御
・スプリットブレイン発生時にサービスを2重起動しない様に、予備系でのサービス起動を抑止

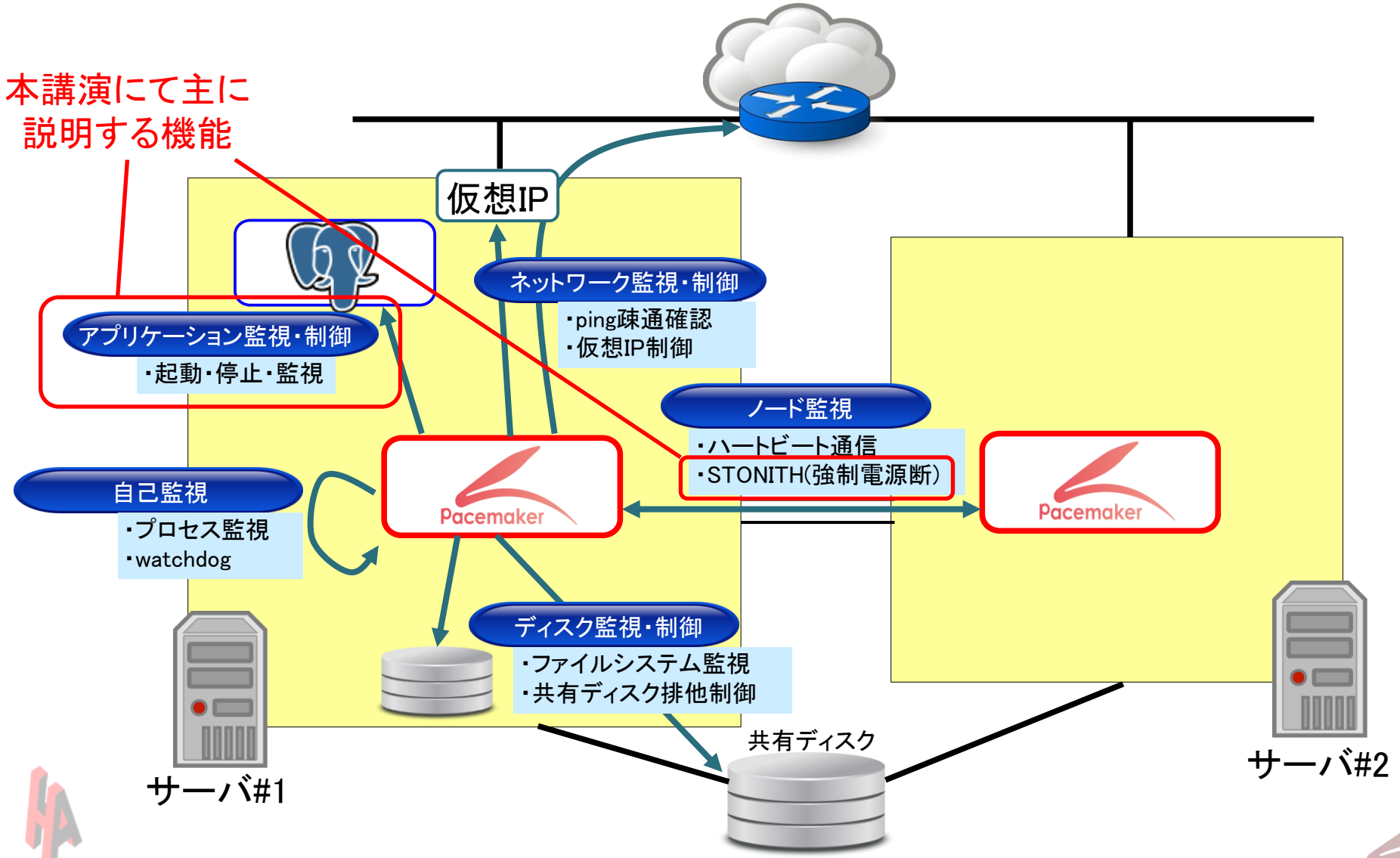
※ スプリットブレインとは
・クラスタ間で正常に通信が行えず、それぞれ独立してサービス管理を行う状態
・サービスが2重起動することにより共有ディスクへの書き込みが競合し、データ破壊が発生するなど、サービス継続に致命的な現象を引き起こす要因となる

＜データ破壊のイメージ＞



Pacemakerにできること

Pacemakerではフェイルオーバや排他制御などを行うために以下の様な機能を実装

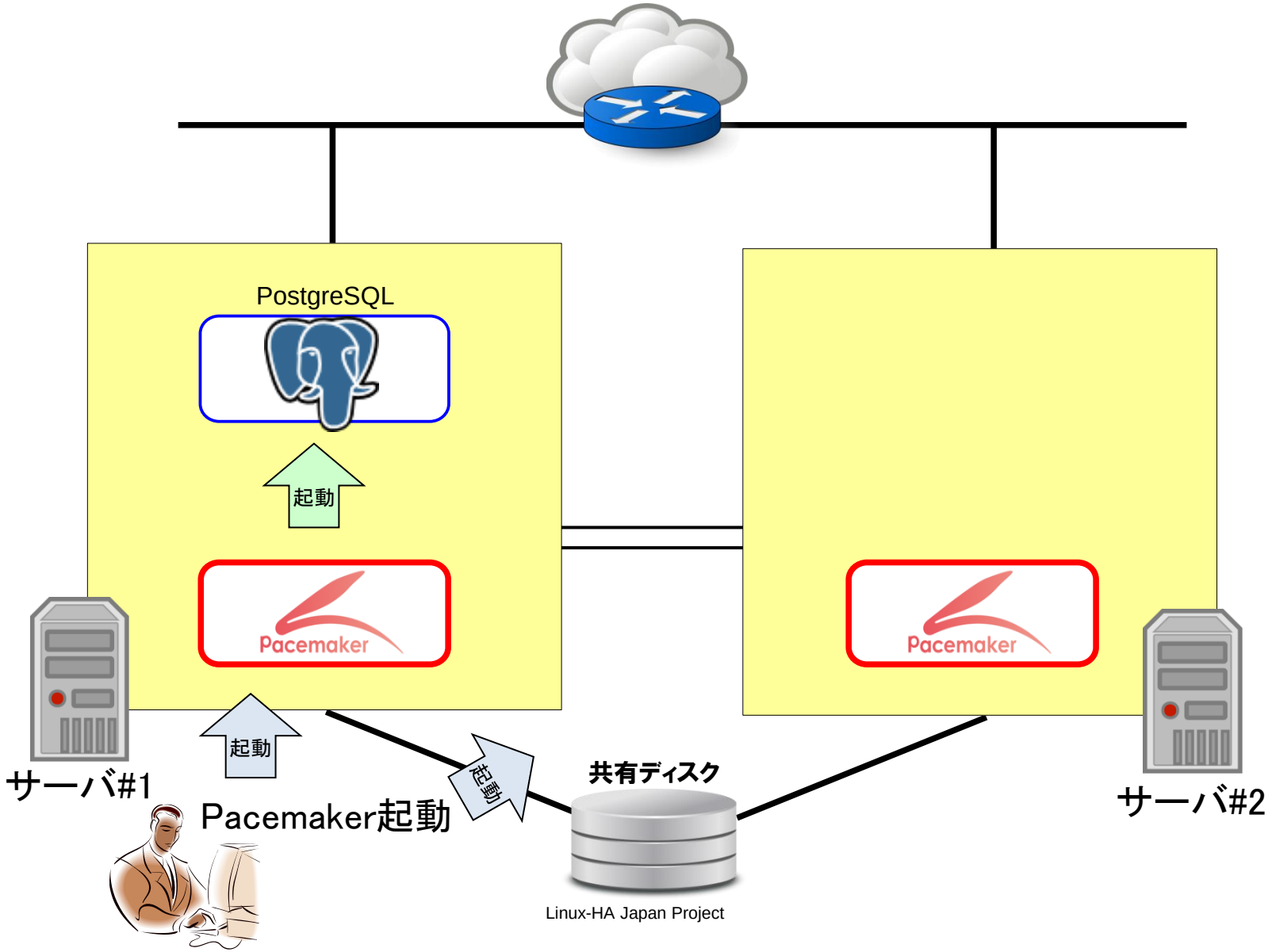


Pacemakerとは

- 故障検知とサービス引継ぎ
- 排他制御

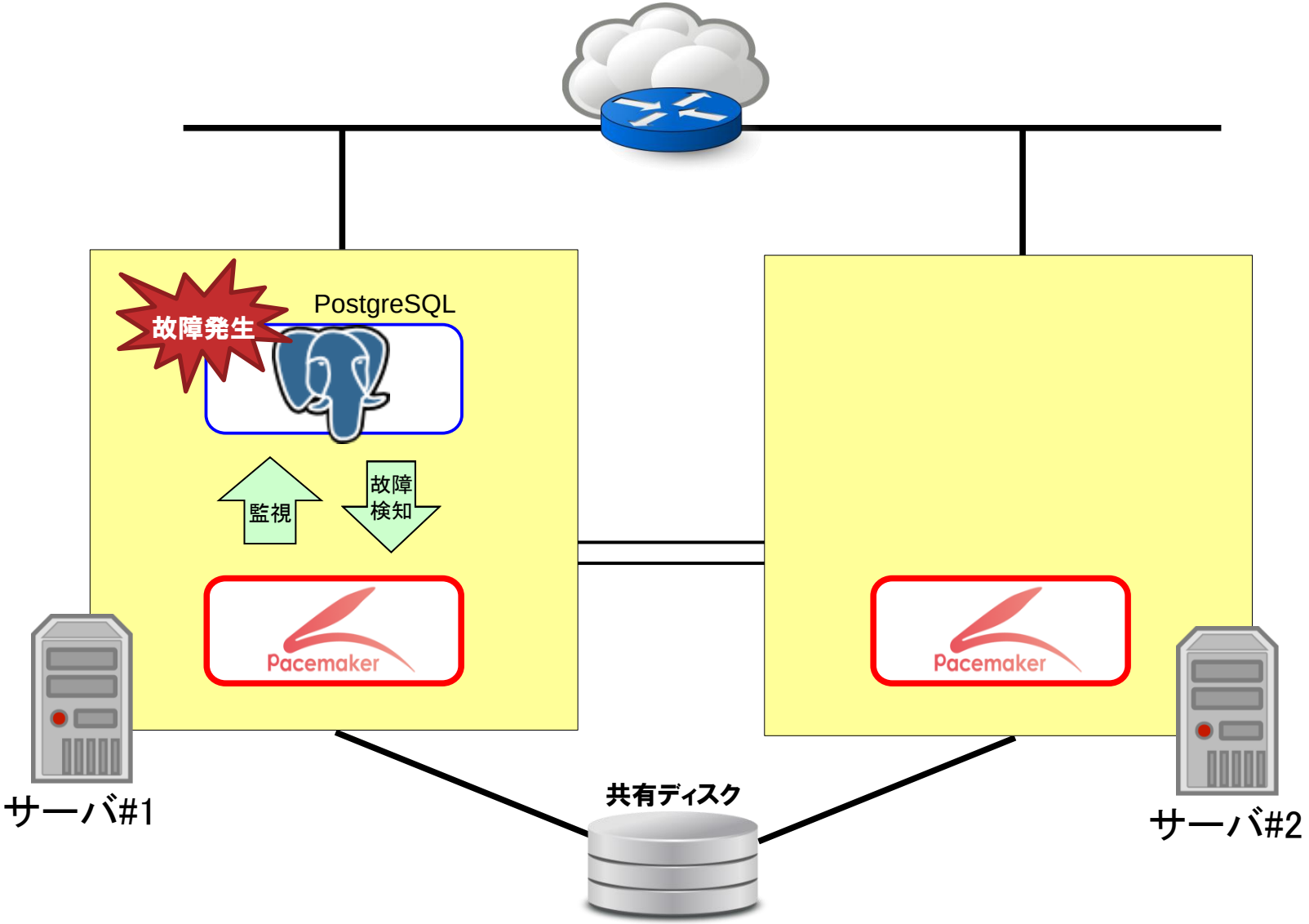
アプリケーションの監視・制御(起動)

アプリケーションなどの起動停止はPacemakerより実施



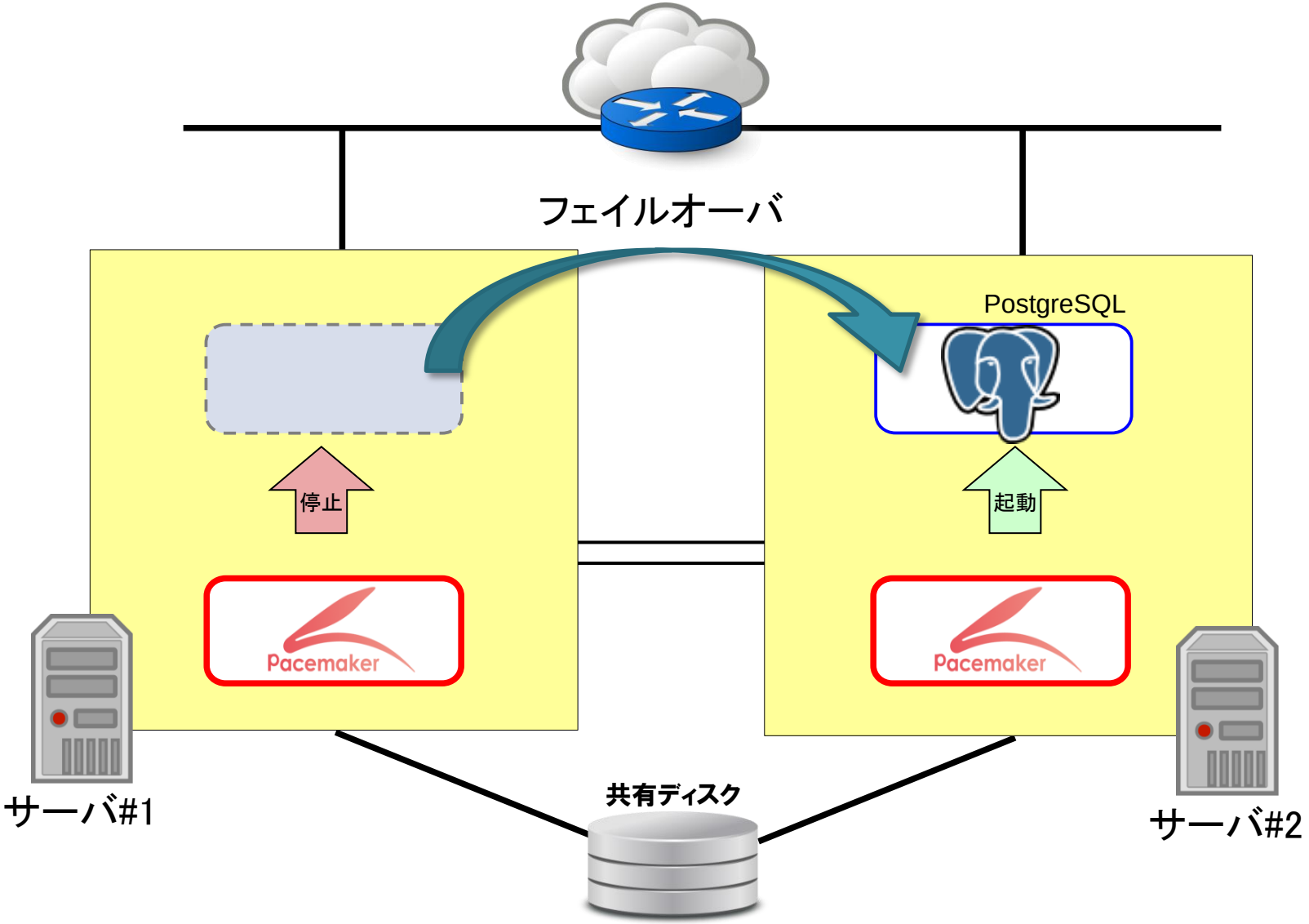
アプリケーションの監視・制御(監視)

Pacemakerはアプリケーション起動後、定期的に監視処理を実施



アプリケーションの監視・制御(フェイルオーバ)

故障検知後はアプリケーションを現用系で停止してから予備系で起動



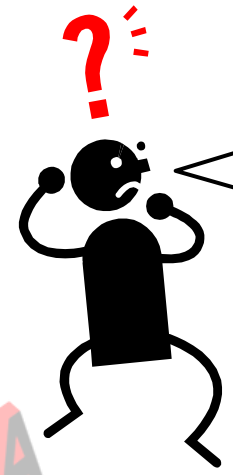
Pacemakerにできること

Pacemakerは様々な対象を起動/停止/監視することが可能

PostgreSQLやApacheの起動停止、Filesystemのmount、仮想IPの割り当てetc....

・・・だけど

どうやって様々な種類のアプリケーション等を監視・制御しているの？



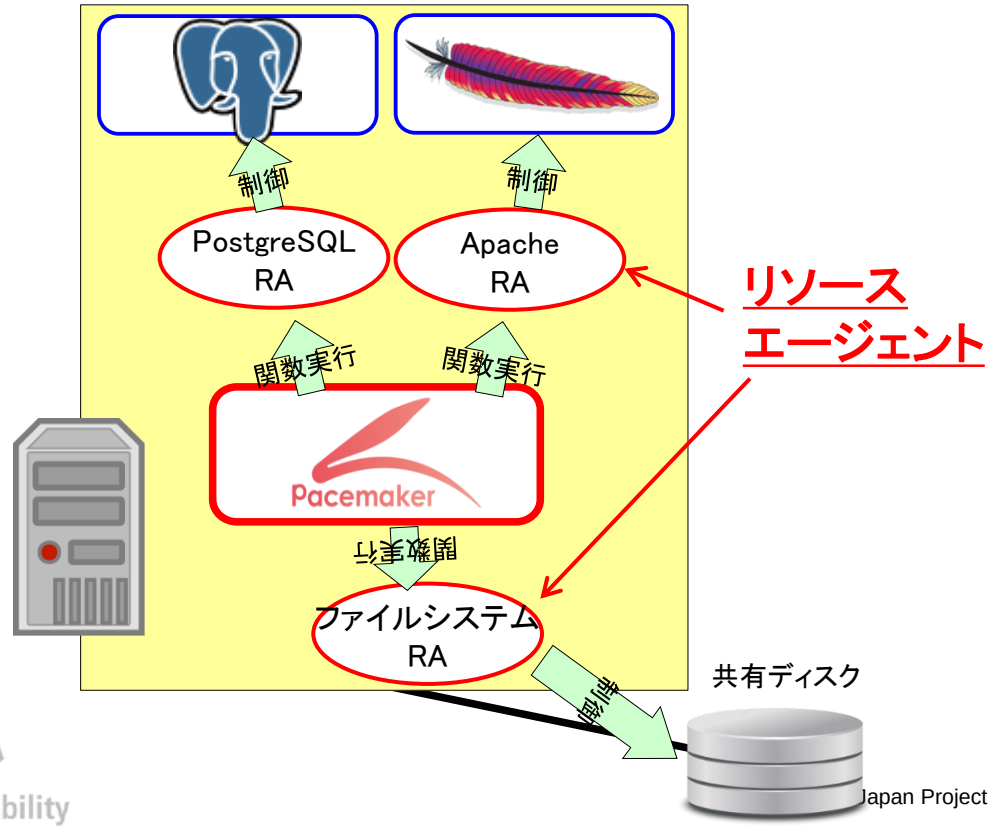
PostgreSQLの起動は　”$ pg_ctl -w start”
Apacheの起動は　”$ apachectl start”

対象毎に起動・停止・監視方法はバラバラなのにどうやって一括管理するの？

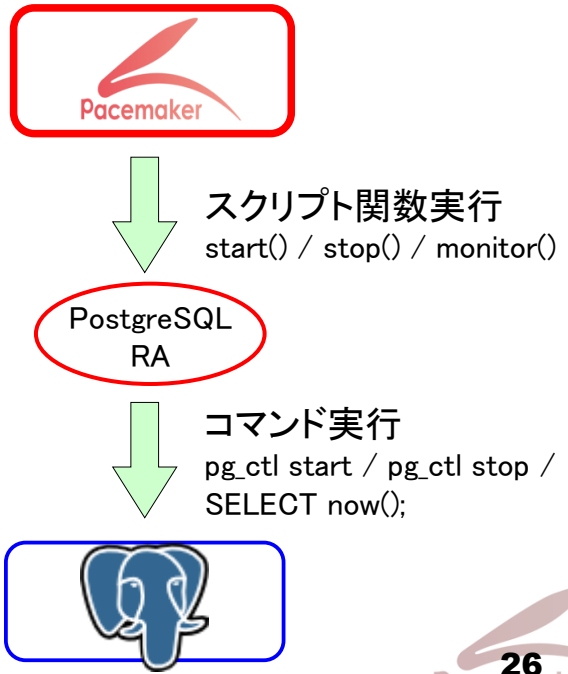
Pacemakerの"故障検知とサービス引継ぎ"の実現

Pacemakerは起動/停止/監視する対象ごとに用意されている「**リソースエージェント(RA)**」と呼ばれる実行ファイルを仲介することで対象にあわせた制御・監視が可能

- 多くのRAはシェルスクリプトで、起動/停止/監視する対象にあわせたstart関数、stop関数、monitor関数などを実装
  (例: PostgreSQL用のRAのstart関数では"$ pg_ctl -w start"を実行)
- Pacemakerではデフォルトで様々な対象ごとのRAを同梱しており、また自作して追加することも可能



PostgreSQLの制御・監視における RA実行の具体的な流れ



Pacemakerに同梱されているRA一覧

以下の様なRAがPacemakerのパッケージに同梱されており、Pacemakerから監視・制御が可能

AoEtarget	LVM-activate		Stateful	aws-vpc-route53		garbd	lxd-info	ovsmonitor
sg_persist	AudibleAlarm		LinuxSCSI	SysInfo	awseip	hulft	machine-info	
pgagent	slapd	CTDB	MailTo	VIPArip	awsvip	iSCSILogicalUnit	mariadb	pgsql
symlink	ClusterMon	ManageRAID		VIPcheck	azure-lb	iSCSITargetminio	pingd	
syslog-ng	Delay	ManageVE	VirtualDomain		clvm	ids	mpathpersist	
portblock	tomcat	Dummy	NodeUtilization		WAS	conntrackd	iface-bridgemysql	
postfix	varnish	EvmsSCC	Pure-FTPd	WAS6	db2	iface-vlan	mysql-proxy	pound
vmware	Evmsd	Raid1	WinPopup	dhcpd	ipsec	nagios	proftpd	vsftpd
Filesystem	Route	Xen	dnsupdate	iscsi	named	rabbitmq-cluster		zabbixserver
ICP	SAPDatabase		Xinetd	docker	jboss	nfsnotify	redis	IPaddr
SAPInstance		ZFS	eDir88	jira	nfsserver	rkt	IPaddr2	SendArp
anything	ethmonitor	kamailio	nginx	rsyncd	IPsrcaddr	ServeRAID	apache	exportfs
ldirectord	oraasm	rsyslog	IPv6addr	SphinxSearchDaemon		asterisk	fio	lvmlockd	oracle
scsi2reservation		LVM	Squid	aws-vpc-move-ip		galera	lxc	oralsnr	sfex
HealthCPU	HealthSMART		SysInfo	SystemHealth		attribute	controld	diskd	
ifspeed	o2cb	ping	remote	などなど….				

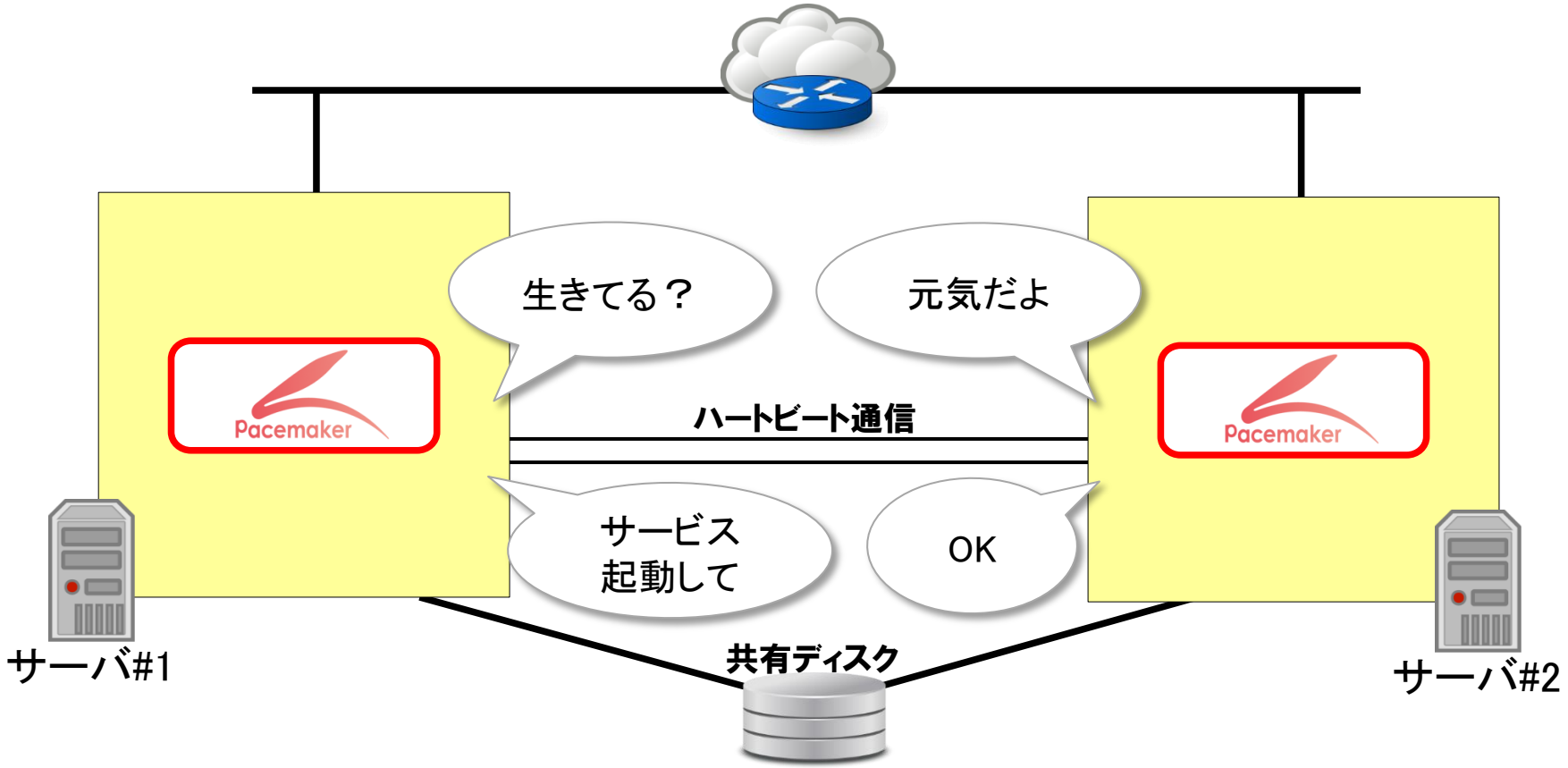
※ systemdのユニットファイルを使用して管理することも可能

Pacemakerとは

- アプリケーションの監視・制御
- **排他制御**

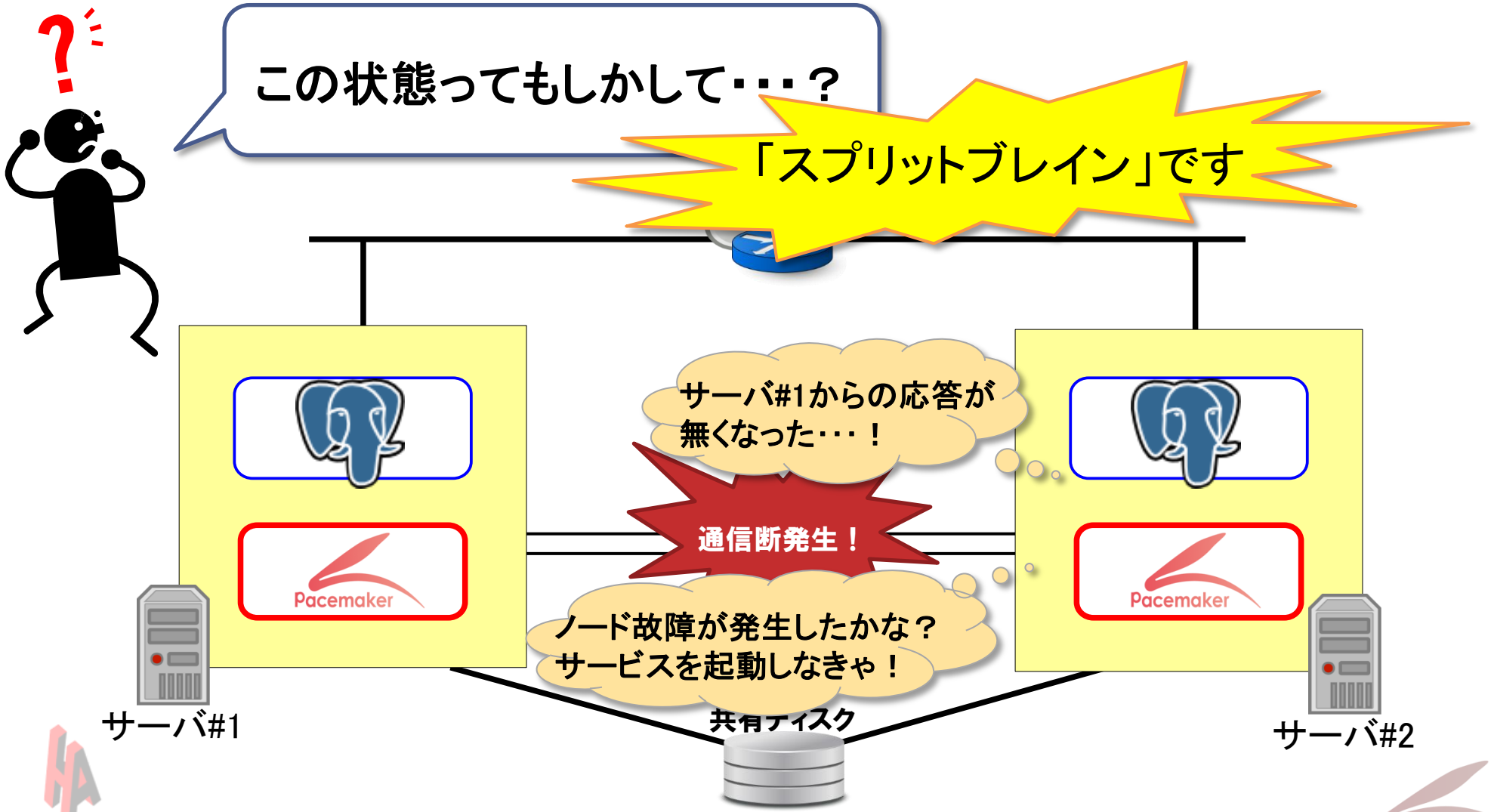
クラスタ間の通信

Pacemakerではクラスタ間で**ハートビート通信**と呼ばれる通信を定期的に送信しあうことでお互いのサーバの状態確認やサービス起動停止の一括制御を実現



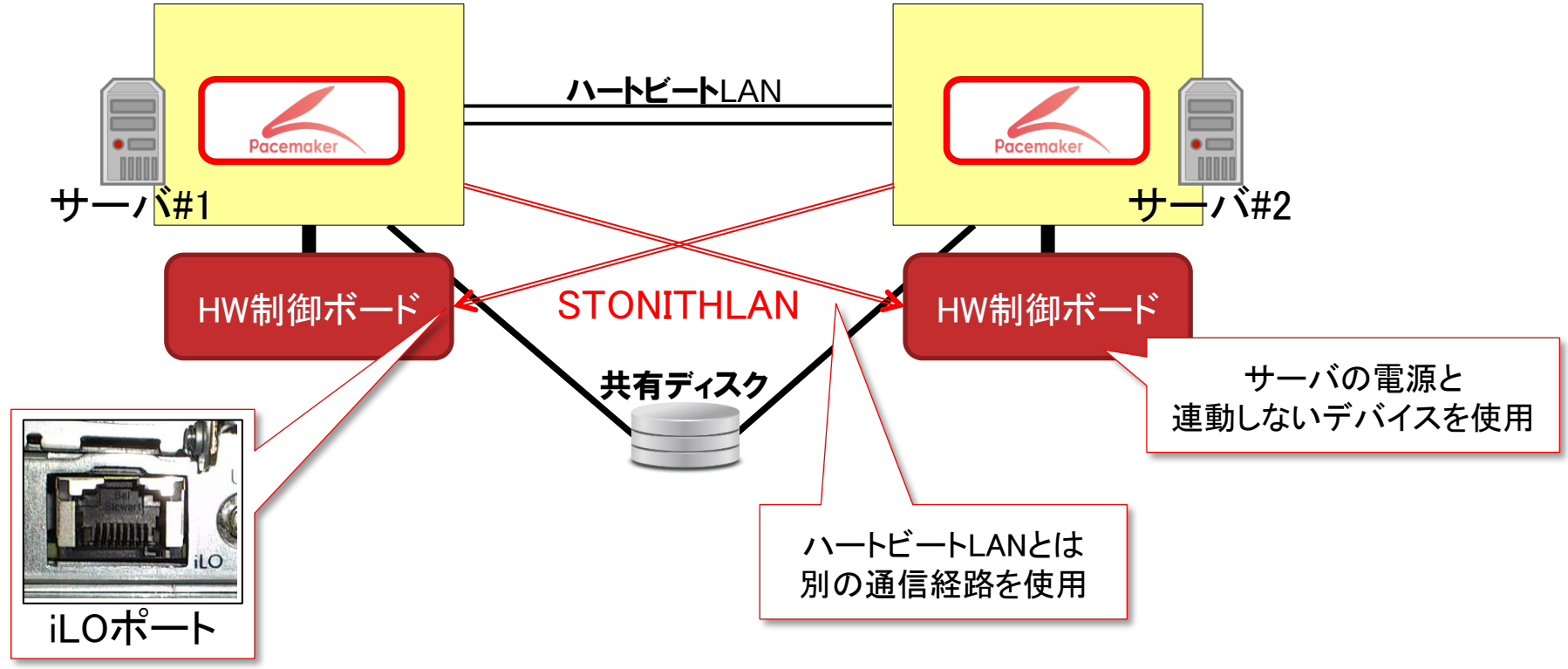
クラスタ間の通信

もしハートビート通信が途絶したら....



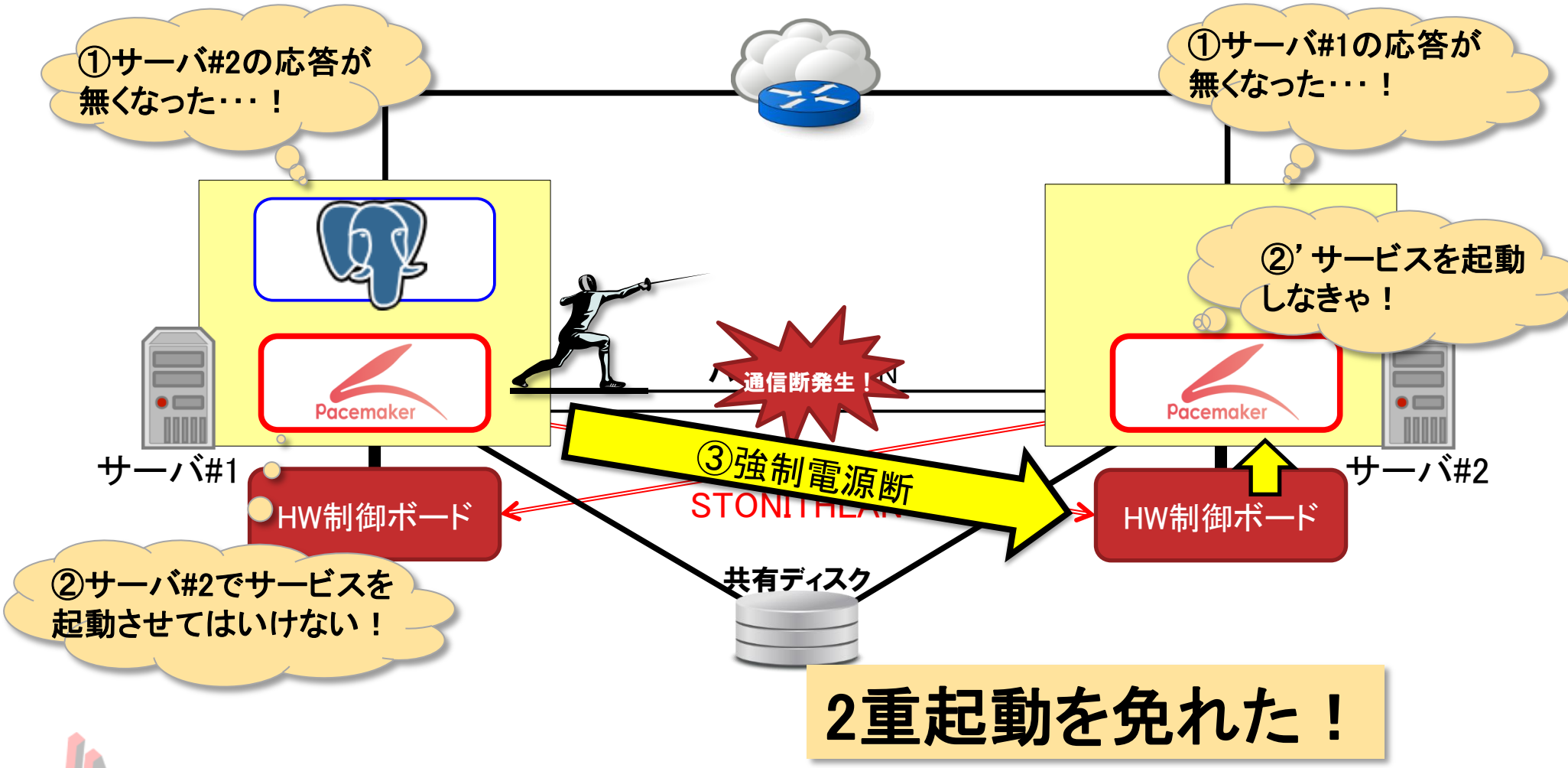
スプリットブレイン対策

- ✓ STONITH(Shoot The Other Node In The Head)
  - ・スプリットブレイン(両系がActive状態)になる前に
    対向ノードの強制電源断を実行する機能(排他制御機能)
  - ・サーバ付属のリモートHW制御ボード(iLOなど)を操作



スプリットブレイン対策

＜STONITHによる排他制御イメージ＞



2重起動を免れた！

【補足】STONITHが使えない環境の場合

HW制御ボードなどが存在せずSTONITHが使用できない場合は
以下のような機能によりスプリットブレイン対策を行うことが可能

- ✓ sfex
  - ・共有ディスクのsfex専用パーティションに、ディスクのロック情報を定期的に書き込む
  - ・Active系によりロック情報が更新されていれば、Active系が生存していると判断し、Standby系でのリソース起動を抑止
  - ・STONITHが使用できる環境においても、信頼性を高めるために本機能を併用することがある

- ✓ VIPcheck
  - ・Standby系からActive系の仮想IP(VIP)に対してpingを送信
  - ・ping応答があれば、Active系が生存していると判断し、Standby系でのリソース起動を抑止

各機能の詳細についてはOSC 2015 Tokyo/Fallに講演している
『試して覚えるPacemaker入門 排他制御機能編』
の資料を参照してください。
➢ http://linux-ha.osdn.jp/wp/archives/4338

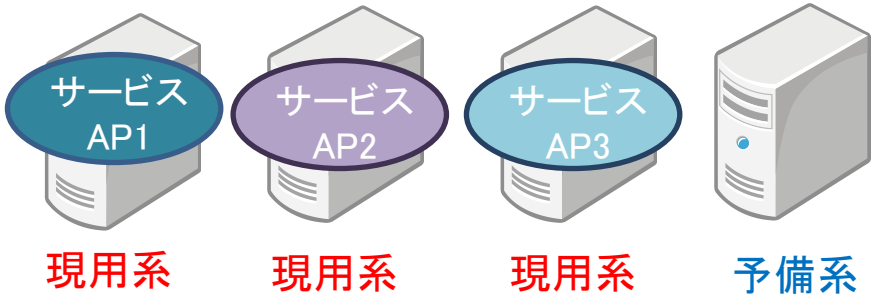
いろんな
HAクラスタ

HAクラスタの構成

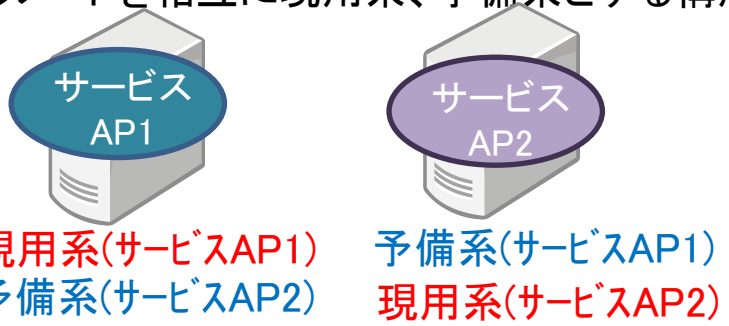
HAクラスタには現用系1台、予備系1台の1+1構成の他に、以下の様な構成も可能

【1+1構成】
現用系と予備系が1:1の構成

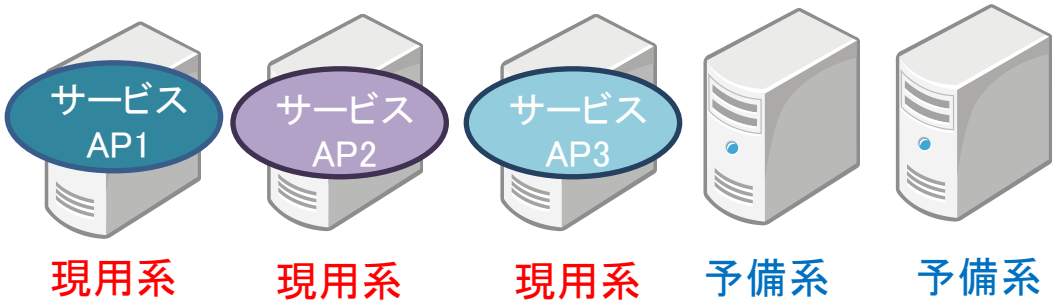
【N+1構成】
複数台の現用系ノードに対して1台の予備系ノードを置く構成



【1+1Cross構成】
2の業務APを別ノードで起動させ、それぞれのノードを相互に現用系、予備系とする構成



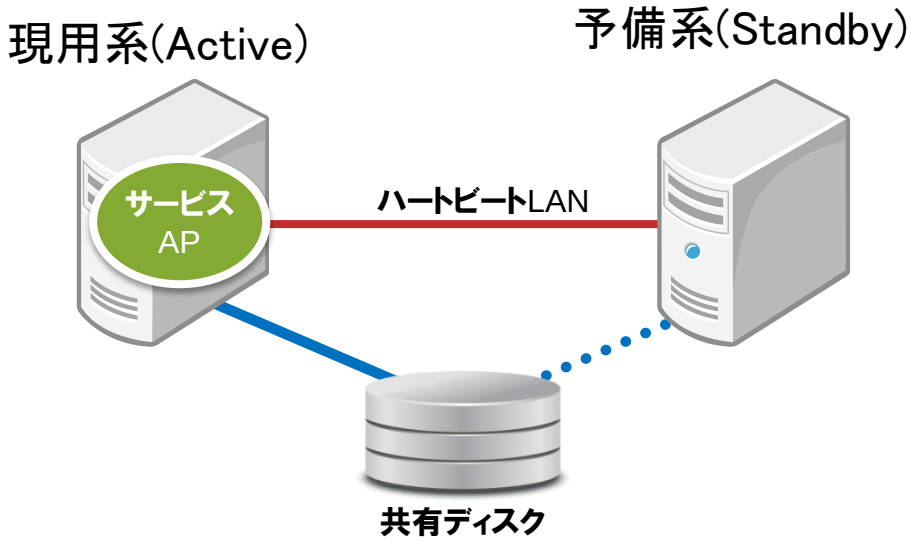
【N+M構成】
複数台の現用系ノードに対して複数台の予備系ノードを置く構成



HAクラスタにおけるデータ管理

HAクラスタではフェイルオーバ後に処理を引き継ぐため、
予備系でも現用系と同一のデータが必要

- 共有ディスクを使用して、同じデータファイルにアクセスする

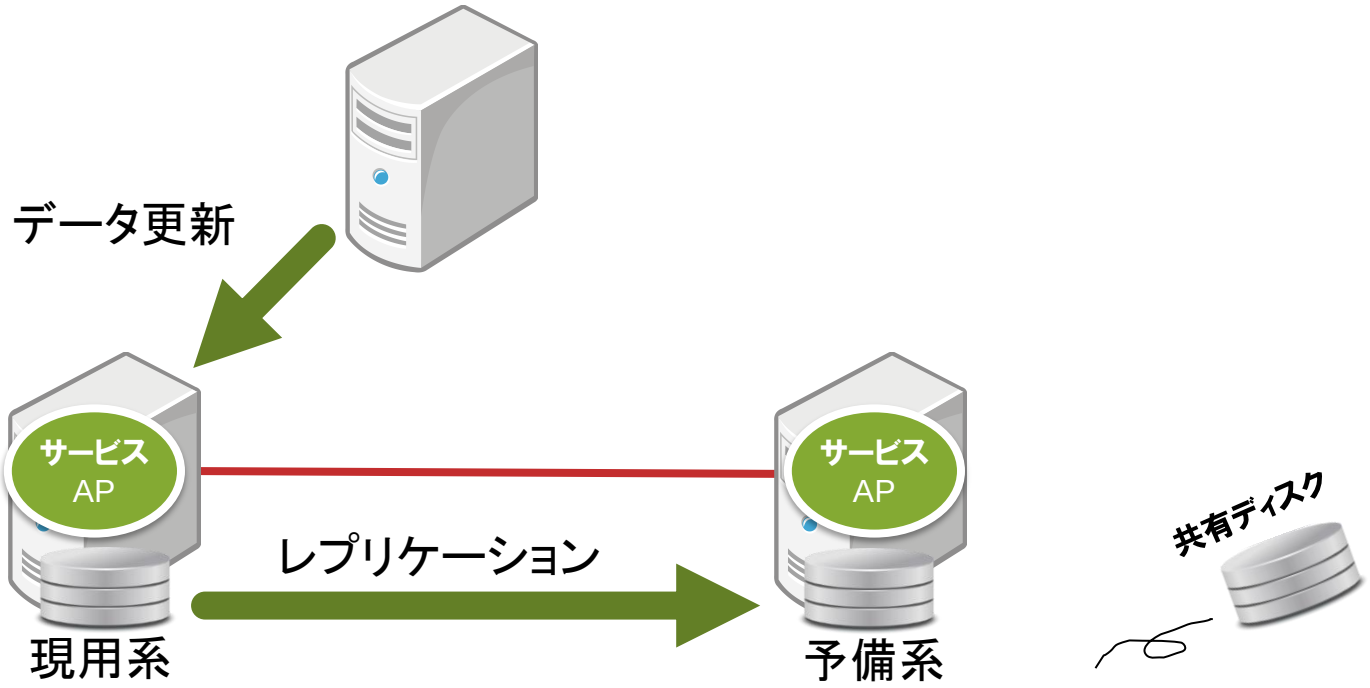


もう一つのやり方が…

- データをレプリケーションして、同じデータを保有する

シェアードナッシングHA構成

データを更新する際に、更新内容をネットワーク経由で予備系にレプリケーションする構成



- OSS製品ではPostgreSQLの機能やDRBDを使用することでレプリケーションが可能。
- 高価な共有ディスクがないためコストは安いがレプリケーションによるオーバヘッドがあるため性能は共有ディスクを使用したほうが高い。

参考資料

PacemakerとPostgreSQLのレプリケーション機能によるシェアードナッシングHA構成についてもっと知りたい方へ

- ✓ オープンソースカンファレンス2017 Kyoto
  「試して覚えるPacemaker入門 PG-REX(Pacemaker + PostgreSQLによるシェアードナッシングHA構成)構築」
  http://linux-ha.osdn.jp/wp/archives/4627
- ✓ オープンソースカンファレンス2018 Osaka
  「試して覚えるPacemaker入門 PG-REX(Pacemaker + PostgreSQLによるシェアードナッシングHA構成)運用」
  http://linux-ha.osdn.jp/wp/archives/4664

PacemakerとDRBDによるシェアードナッシングHA構成についてもっと知りたい方へ

- ✓ SIOS社(旧サードウェア社)提供の日本語マニュアル
  https://blog.3ware.co.jp/ドキュメント/

まとめ

まとめ

- ✓ 可用性とクラスタ
  - 可用性を向上させるためにクラスタを導入しSPOFを除去
  - 「負荷分散クラスタ」の同一のサービスを複数サーバで提供することで処理の分散と可用性の向上を実現
  - 「HAクラスタ」は故障発生時に同一のサービスを別のサーバで立ち上げなおすことで可用性の向上を実現
  - 「負荷分散クラスタ」と「HAクラスタ」はサービスを2重起動するか否かでどちらの構成とするか決める

- ✓ PacemakerによるHAクラスタ
  - HAクラスタソフトの主な機能として故障検知とサービス引継ぎ、スプリットブレイン対策のための排他制御がある
  - Pacemakerではリソースエージェント(RA)を使用することで様々な対象の故障検知やサービス引継ぎが可能
  - スプリットブレイン対策としてSTONITHが使用可能
  - HAクラスタは現用系、予備系を1+1構成だけではなくN+M構成など様々な構成が可能
  - 動的なデータは共有ディスクの使用やレプリケーションすることで予備系に共有

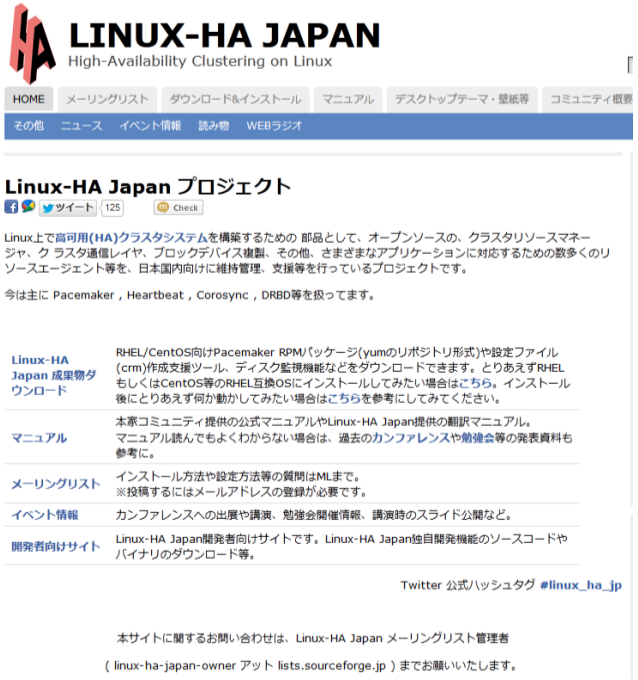
コミュニティ紹介

コミュニティ紹介

Linux-HA Japan URL

http://linux-ha.osdn.jp/

https://ja.osdn.net/projects/linux-ha/



Pacemaker関連の最新情報を日本語で発信

Pacemakerのダウンロードもこちらからどうぞ
(インストールが楽なリポジトリパッケージを公開しています)

コミュニティ紹介

日本におけるHAクラスタについての活発な意見交換の場として
「Linux-HA Japan日本語メーリングリスト」も開設しています

Linux-HA-Japan MLでは、Pacemaker、Heartbeat3、Corosync
DRBDなど、HAクラスタに関連する話題は歓迎！

- ML登録用URL
  http://linux-ha.osdn.jp/
  の「メーリングリスト」をクリック

- MLアドレス

linux-ha-japan@lists.osdn.me

※スパム防止のために、登録者以外の投稿は許可制です

ご清聴ありがとうございました