

Haute disponibilité appliquée aux machines virtuelles

1. vSphere HA

a. Introduction

vSphere HA utilise les ressources d'un serveur vCenter et d'un cluster vSphere pour offrir une haute disponibilité des machines virtuelles et de leurs applications.

vSphere HA propose les fonctionnalités suivantes :

- Protection contre la panne d'un serveur en redémarrant les machines virtuelles sur d'autres hôtes.
- Protection contre la panne système ou applicative en surveillant constamment la machine virtuelle et en la redémarrant en cas de plantage.
- Protection contre les problèmes d'accès de banque de données (*datastore*).
- Protection contre l'isolement réseau d'un hôte, auquel cas les machines virtuelles peuvent être redémarrées sur d'autres hôtes non isolés.

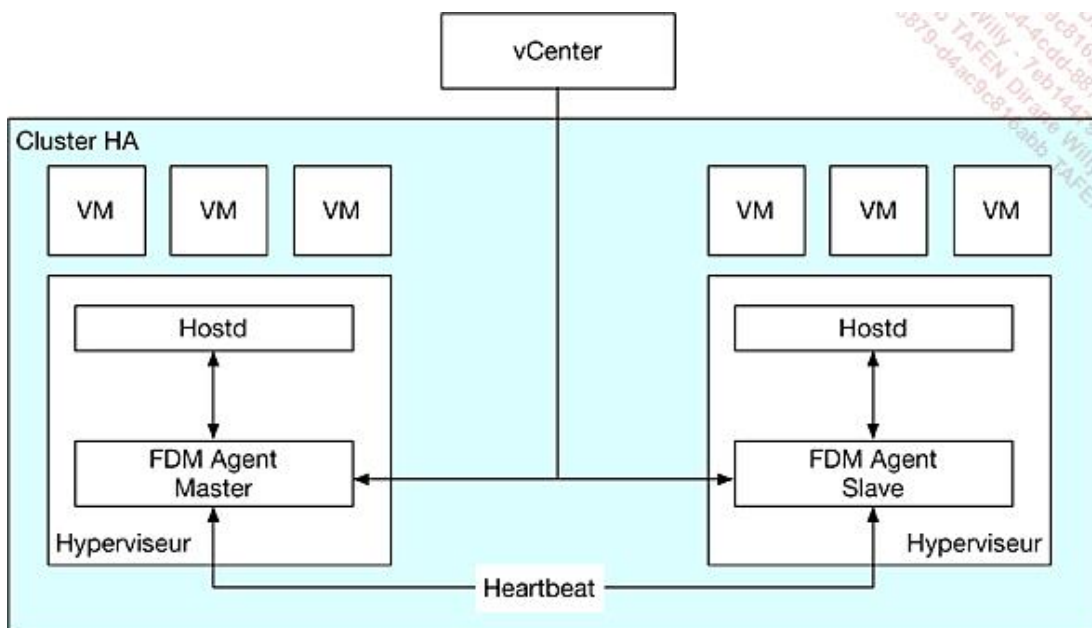
Il est à noter que HA fonctionnant au niveau des hyperviseurs, il n'est pas nécessaire d'installer un agent HA au sein de chaque système invité (machine virtuelle) s'exécutant au sein de notre architecture de virtualisation. Cependant, notez que vSphere HA utilise les VMware Tools pour certaines de ses opérations (surveillances des VMs).

Par ailleurs, on dit des machines virtuelles sur lesquelles HA opère une surveillance qu'elles sont **protégées**.

Décrivons l'architecture de vSphere HA.

b. Architecture de vSphere HA

Le schéma suivant décrit l'architecture de la solution vSphere HA :



Un cluster HA comporte au maximum 64 serveurs ESXi.

Un agent FDM (*Fault Domain Manager*) est exécuté à la fois sur le maître et les esclaves. Celui-ci est notamment en charge d'envoyer à intervalle régulier un signal de pulsation (*heartbeat*) pour s'assurer du statut des hôtes. Les agents communiquent via le(s) réseau(x) de gestion des serveurs ESXi.

Examinons maintenant comment fonctionne l'initialisation de ces composants sur les hôtes et le mécanisme d'élection.

c. Initialisation et élection du maître HA et des esclaves

À l'initialisation d'un cluster HA ou à l'ajout d'un hôte dans un cluster HA, un agent HA est automatiquement installé sur les hôtes. L'agent HA est configuré sur chaque hôte (et automatiquement) pour communiquer avec les autres agents des membres du cluster.

Chaque hôte participant à un cluster vSphere HA aura un rôle attribué, maître (« Master ») ou esclave (« Slave »). Un cluster contient un nœud maître quand les autres membres occuperont le rôle d'esclave.

Le rôle de chaque hôte est déterminé à l'aide d'une élection. Un hôte disposant d'un nombre plus important de banques de données (*datastore*) montées aura un avantage lors de celle-ci.

Si l'administrateur le souhaite, il est possible de forcer la configuration d'un maître spécifique en modifiant le paramètre avancé de l'hôte `fdm.nodeGoodness` et en lui assignant une valeur importante (comme par exemple 1000), la plus grande valeur permettant à l'hôte de gagner l'élection et donc de devenir maître du cluster.

Si deux serveurs ont le même nombre de *datastore*, c'est le serveur ayant le plus grand MOID (*Managed Object ID*) qui est choisi, référence unique attribuée aux hôtes par vCenter. Attention, il s'agit ici d'une comparaison lexicale des chiffres, ce qui veut dire que 99 sera considéré comme plus grand que 100 car 9 est supérieur lexicalement à 1.

Le maître d'un cluster HA possède les responsabilités et rôles suivants :

- Supervise l'état des hôtes esclaves. C'est le maître en cas de panne d'un des hôtes esclaves qui détermine les machines virtuelles à redémarrer.
- Surveille l'état des machines virtuelles protégées par HA. Si une machine virtuelle venait à être inopérante, le maître prend la décision de redémarrer celle-ci (et également sur quelle ressource le redémarrage aura lieu).
- Maintient une vue complète sur les machines et hôtes protégés
- Agit comme interface de gestion de la perspective de vCenter.
- Signale le statut du cluster HA à vCenter

Le rôle des hôtes esclaves (de la perspective de HA) se résume à exécuter des machines virtuelles et à en reporter le statut au maître. En outre, le maître peut également exécuter des machines virtuelles, en plus de ses attributions.

Dans le cas où le maître serait indisponible, une nouvelle élection aurait lieu pour désigner un nouveau maître du cluster HA. Une nouvelle élection est déclenchée dans les scénarios suivants :

- Le maître est défaillant.
- Le réseau est partitionné ou isolé.
- Le maître est déconnecté du serveur vCenter.
- Le maître est mis en état de maintenance ou en standby.

- Lorsque vSphere HA est reconfiguré.

Le maître crée et maintient un fichier nommé `protectedlist`. Ce fichier décrit les machines virtuelles protégées par HA et leurs états. Il est stocké sur les banques de données configurées pour participer au cluster HA dans le cadre du *datastore heartbeat*. Ce fichier est stocké dans le répertoire caché dédié à vSphere HA :

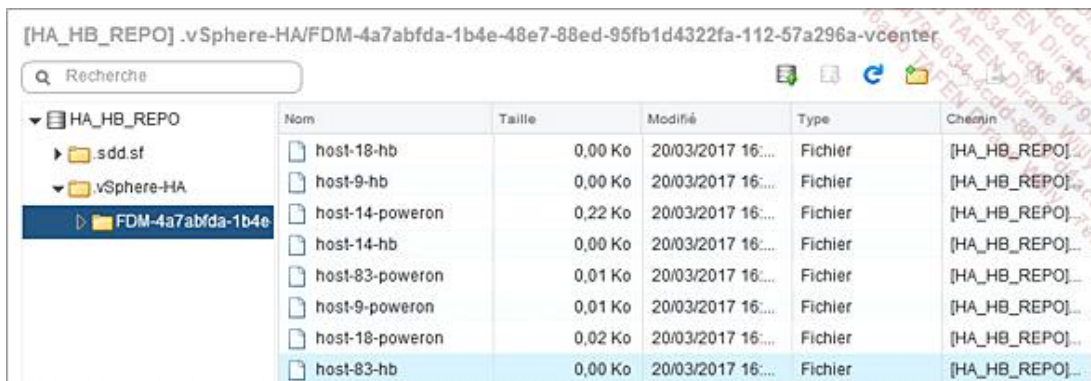
```
<datastore>/ .vSphere-HA/<repertoire-cluster>/protectedlist
```

Les valeurs ci-dessus en gras changent avec la racine de la banque de données et le répertoire dédié au cluster.

Dans ce même répertoire caché, il existe un fichier présent sur l'ensemble des banques de données configurées pour contribuer à un cluster HA :

- Le fichier `host-<host-number>-poweron` qui permet de garder l'information sur l'état des machines virtuelles (allumées ou éteintes) sur chaque serveur hôte.
Il sert aussi de marqueur afin de déterminer l'état des serveurs. Le serveur maître lit le fichier pour déterminer si un serveur esclave est isolé du reste du réseau. Si le maître trouve une valeur égale à 0, le serveur esclave portant le fichier n'est pas isolé. Dans le cas où ce dernier trouve une valeur égale à 1, le serveur s'est signalé comme isolé en changeant cette valeur dans ce fichier.
- Le fichier `host-<host-number>-hb` sont les fichiers utilisés dans le cadre du *datastore heartbeat*. Ils sont mis à jour régulièrement pour permettre de déterminer si l'hôte est toujours « vivant » de la perspective du *datastore*.

La capture ci-dessous montre ces fichiers dans l'une des banques de données utilisées pour le *datastore heartbeat*.



Nom	Taille	Modifié	Type	Chemin
host-18-hb	0,00 Ko	20/03/2017 16:...	Fichier	[HA_HB_REPO]...
host-9-hb	0,00 Ko	20/03/2017 16:...	Fichier	[HA_HB_REPO]...
host-14-poweron	0,22 Ko	20/03/2017 16:...	Fichier	[HA_HB_REPO]...
host-14-hb	0,00 Ko	20/03/2017 16:...	Fichier	[HA_HB_REPO]...
host-83-poweron	0,01 Ko	20/03/2017 16:...	Fichier	[HA_HB_REPO]...
host-9-poweron	0,01 Ko	20/03/2017 16:...	Fichier	[HA_HB_REPO]...
host-18-poweron	0,02 Ko	20/03/2017 16:...	Fichier	[HA_HB_REPO]...
host-83-hb	0,00 Ko	20/03/2017 16:...	Fichier	[HA_HB_REPO]...

Par ailleurs, le journal de l'agent FDM de chaque hôte peut être consulté à l'aide du fichier `/var/log/fdm.log`.

```

[root@esx1:~] tail -n 5 -f /var/log/fdm.log
2017-03-18T16:13:45.933Z info fdm[2FE0B70] [Originator@6876 sub=Cluster opID=SWI
-3ab50c2a] [ClusterManagerImpl::LogState] hostId=host-9 state=Master master=None
isolated=false host-list-version=8 config-version=6 vm-metadata-version=2 slv-m
st-tdiff-sec=0
2017-03-18T16:23:46.558Z info fdm[2FE0B70] [Originator@6876 sub=Cluster opID=SWI
-3ab50c2a] [ClusterManagerImpl::LogState] hostId=host-9 state=Master master=None
isolated=false host-list-version=8 config-version=6 vm-metadata-version=2 slv-m
st-tdiff-sec=0
2017-03-18T16:33:47.272Z info fdm[2FE0B70] [Originator@6876 sub=Cluster opID=SWI
-3ab50c2a] [ClusterManagerImpl::LogState] hostId=host-9 state=Master master=None
isolated=false host-list-version=8 config-version=6 vm-metadata-version=2 slv-m
st-tdiff-sec=0
2017-03-18T16:43:48.089Z info fdm[2FE0B70] [Originator@6876 sub=Cluster opID=SWI
-3ab50c2a] [ClusterManagerImpl::LogState] hostId=host-9 state=Master master=None
isolated=false host-list-version=8 config-version=6 vm-metadata-version=2 slv-m
st-tdiff-sec=0
2017-03-18T16:53:48.925Z info fdm[2FE0B70] [Originator@6876 sub=Cluster opID=SWI
-3ab50c2a] [ClusterManagerImpl::LogState] hostId=host-9 state=Master master=None
isolated=false host-list-version=8 config-version=6 vm-metadata-version=2 slv-m
st-tdiff-sec=0

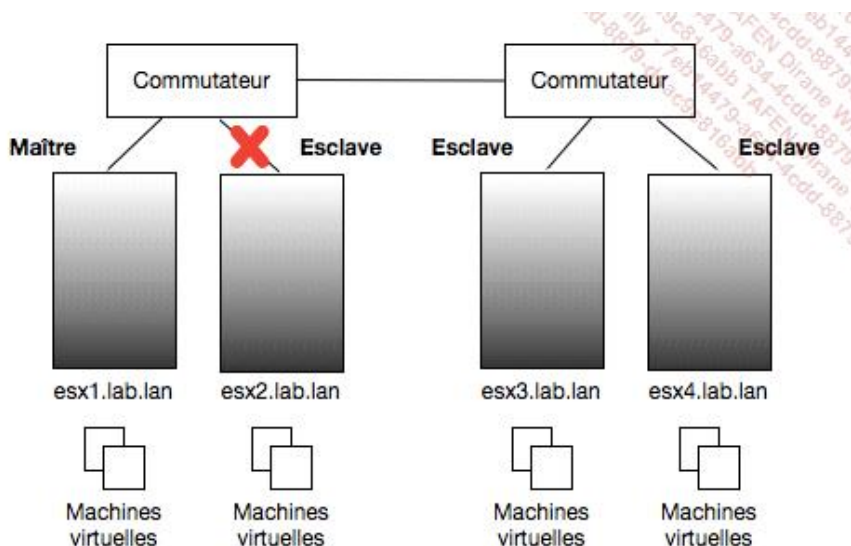
```

d. Heartbeat, types de pannes et de réponses vSphere HA

HA valide l'état de disponibilité des serveurs vSphere via un mécanisme nommé heartbeat (signal de pulsation). L'utilisation de ce mécanisme et de la configuration du cluster permet de déterminer la réaction de HA en cas d'indisponibilité d'un serveur.

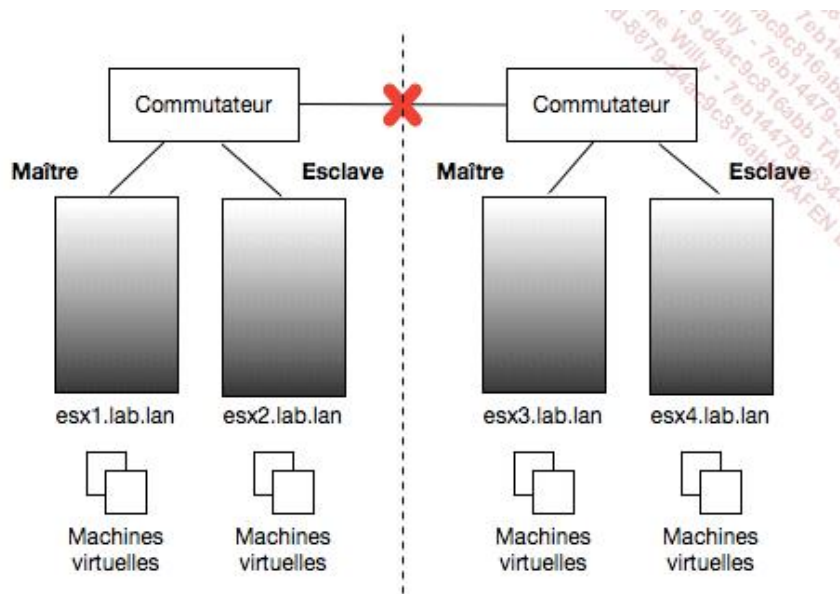
Trois types de panne sont détectées par vSphere HA, à savoir :

- **Panne (Failure)** : un hôte ne fonctionne plus correctement.
- **Isolement** : un hôte fonctionne toujours mais devient isolé par rapport au reste du réseau.



Cela peut se produire en cas de la panne d'un adaptateur réseau ou d'un lien entre l'hyperviseur et le commutateur.

- **Partition** : un hôte fonctionne toujours mais perd la connectivité avec l'hôte maître.



La partition est ici générée par le fait que le lien intercommutateur qui permet la communication de l'ensemble du cluster HA n'est pas fonctionnel. Dans le cas d'une partition, les hôtes sur les deux segments partitionnés auront comme réaction d'élire sur chaque zone isolée, un maître HA.

Le maître du cluster HA surveille en permanence le statut des hôtes via le réseau à l'aide de heartbeats et plus particulièrement par le biais de requêtes ICMP (pings) envoyés toutes les secondes aux interfaces de gestion des esclaves.

Dans le cas d'une non-réponse, le maître décide de vérifier si l'hôte ne serait pas « vivant » de la perspective d'une banque de données avec qui l'hôte est censé communiquer (*datastore heartbeat*).

Si cet essai se révèle être positif, le maître considère que l'hôte est en état d'**isolement** ou de **partition** et continue de surveiller l'hôte et ses machines virtuelles.

Dans le cas d'un **isolement**, lorsqu'un hôte ne voit plus de trafic émanant de la part d'autres agents HA, il essaie de joindre l'adresse d'isolement (ou les adresses d'isolement) du cluster. Si cela ne fonctionne pas, il se considère comme isolé du reste du réseau et reflète cet état dans les datastores utilisés pour le heartbeating. À cette occasion, une réponse (paramétrée par l'administrateur) peut être de redémarrer les machines virtuelles sur un autre hôte non isolé.

Une adresse d'isolement est une adresse IP que les serveurs ESXi essaient de joindre dans le cas où ils ne voient pas passer du trafic de heartbeat sur les interfaces de management.

C'est par défaut l'adresse de passerelle du réseau de management. Il est néanmoins possible de créer plusieurs adresses d'isolement via le paramètre avancé du cluster `das.isolationaddressX` où X correspond à un chiffre/nombre entre 1 et 10 avec la valeur [adresse IP].

Vous pouvez définir donc jusqu'à 10 adresses d'isolement. Dans ce cas, il convient d'indiquer au cluster d'utiliser les adresses définies comme adresses d'isolement via le paramètre `das.usedefaultisolationaddress = false`.

Si le dernier paramètre n'est pas configuré, le cluster ignorera les adresses d'isolation et utilisera la passerelle du réseau de management. Cela pourrait être gênant si la passerelle est par exemple un switch cœur de réseau ou un pare-feu vers lequel on interdit la réponse à la commande ping.

Si cet essai est infructueux, l'hôte est déclaré comme en **panne** et les machines virtuelles peuvent être redémarrées sur d'autres membres du cluster.

e. Proactive HA

vSphere HA peut également être capable de réagir en cas de panne non vitale, c'est ce que l'on appelle le Proactive HA.

Proactive HA est capable de surveiller le statut des composants matériels d'un serveur. Ainsi, en cas de pertes d'un des composants d'un hyperviseur (alimentation, réseau...), il peut être décidé en anticipation d'évacuer les machines virtuelles vers d'autres hyperviseurs, par le biais de vMotion. Et ce, même si les machines virtuelles n'ont pas été impactées par la panne d'un des composants de l'hôte.



En complément de l'activation de cette fonctionnalité, notez que vous devez activer dans votre cluster la répartition de charge DRS pour pouvoir bénéficier de ce mécanisme.

f. Haute disponibilité système et applicative

Il est difficile de juger de l'état opérationnel d'une machine virtuelle en se basant uniquement sur son statut d'alimentation. vSphere HA propose aux administrateurs d'aller au-delà de cette simple surveillance.

Ainsi, il est possible de s'assurer du fonctionnement nominal du système d'exploitation virtualisé mais également d'une application s'exécutant à l'intérieur d'une machine virtuelle, grâce aux VMware Tools.

Le fonctionnement nominal d'une machine virtuelle est évalué à l'aide de *heartbeats* envoyés à intervalle régulier par les VMware Tools s'exécutant sur le système invité et par son activité système (activité disque/activité réseau).

Le statut d'une application peut également être surveillé par le cluster (ApplicationHA) si les développeurs ont intégré à celle-ci le SDK fourni par VMware. Des éditeurs comme Symantec utilisent ce mécanisme pour assurer une haute disponibilité applicative.

Dans le cas où le système d'exploitation venait à dysfonctionner, le cluster peut automatiquement prendre la décision de redémarrer la machine virtuelle. Cette décision aboutit après un processus précis, à savoir :

- Le maître du cluster vérifie si un ou plusieurs *heartbeats* émanant de VMware Tools sont détectés dans l'intervalle de panne configuré.
- En cas de non-réception de *heartbeats* dans cet intervalle, le maître du cluster évalue, par défaut pendant 120 secondes, une quelconque activité disque ou réseau. Cela évite de redémarrer une machine virtuelle toujours fonctionnelle mais qui n'envoie plus de *heartbeats*, ce qui peut se produire si les VMware Tools ne s'exécutent plus sur la machine pour une quelconque raison.
- Si aucune activité n'est détectée durant ce laps de temps, la machine est redémarrée.

La surveillance applicative fonctionne de la même façon. Dans le cas où les *heartbeats* applicatifs ne seraient pas reçus dans un intervalle de temps configuré par l'administrateur, la machine virtuelle serait redémarrée.

Il est nécessaire d'ajouter que pour éviter de faire redémarrer une machine virtuelle en boucle, ce processus se répétera par défaut 3 fois au maximum dans la période de redémarrage configurée, selon la sensibilité choisie.

De plus, sachez que dans le cadre du démarrage d'une machine virtuelle protégée par vSphere HA, le maître du cluster ne surveillera pas la machine virtuelle avant un laps de temps, spécifié par le paramètre de temps de disponibilité minimal, cela permet d'éviter les faux-positifs.

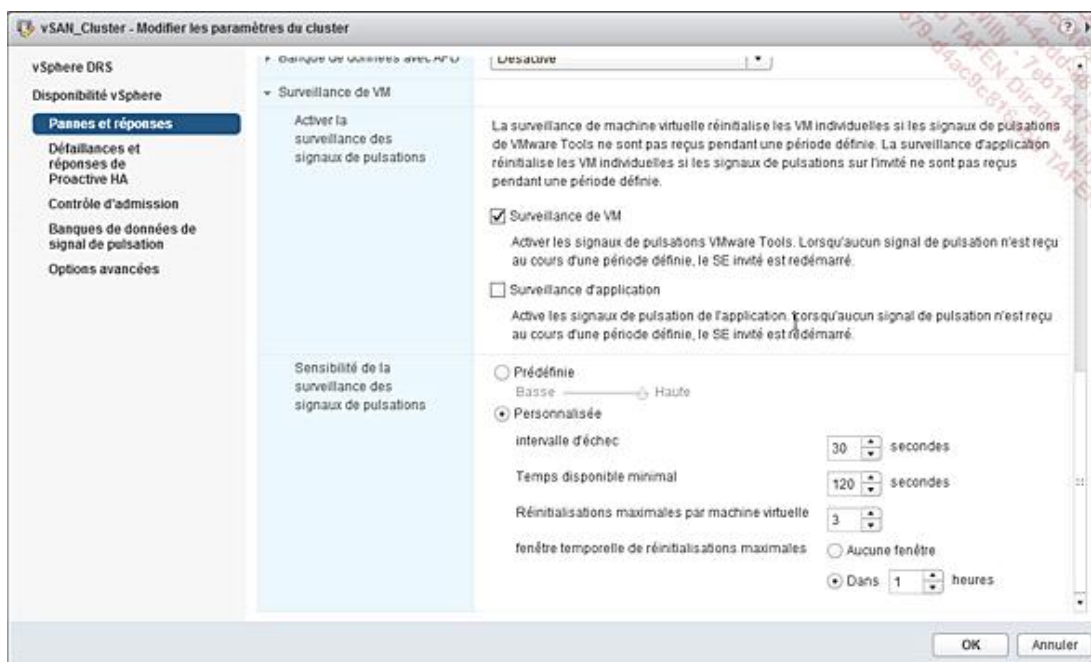
Le tableau ci-dessous illustre les différentes sensibilités disponibles et leurs paramètres par défaut :

Sensibilité	Intervalle de panne (en secondes)	Temps de disponibilité minimal (en secondes)	Période de redémarrage
Haute	30	120	1 heure
Moyenne	60	240	24 heures
Basse	120	480	7 jours

Comme vous pouvez le constater, plus la sensibilité est haute, plus le nombre potentiel de redémarrages pouvant intervenir dans un laps de temps est augmenté. Si l'administrateur choisit la sensibilité haute, ce dernier autorisera une machine virtuelle à être redémarrée automatiquement trois fois au maximum dans l'heure.

En outre, notez que l'ensemble de ces paramètres (intervalle de panne, temps de disponibilité minimal, période de redémarrage, nombre de redémarrages maximum) peuvent être manuellement spécifiés par l'administrateur, hors utilisation des 3 niveaux de sensibilité définis par défaut dans le cluster.

La figure ci-dessous montre ces paramètres modifiables :



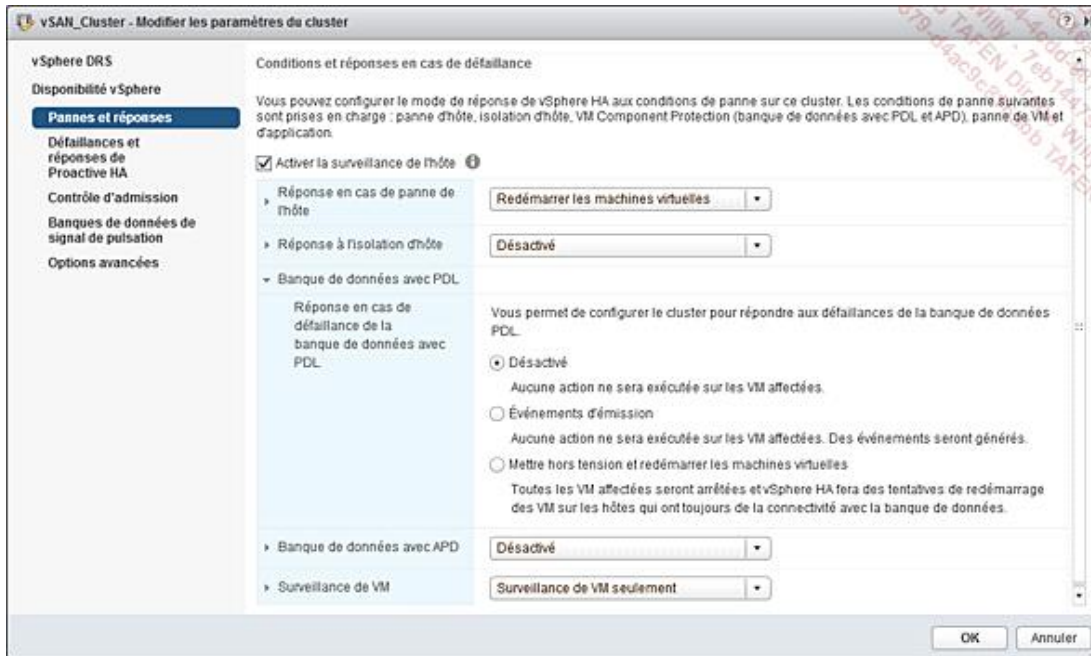
g. VM Component Protection

VM Component Protection ou VMCP est une nouvelle fonctionnalité de vSphere 6.0. Elle fournit une protection contre la perte d'accès aux banques de données pouvant affecter une machine virtuelle dans un cluster vSphere HA. Lorsqu'un problème d'accès au datastore se présente, un serveur vSphere n'a plus accès au stockage impacté.

Il existe deux types de perte d'accès aux datastores :

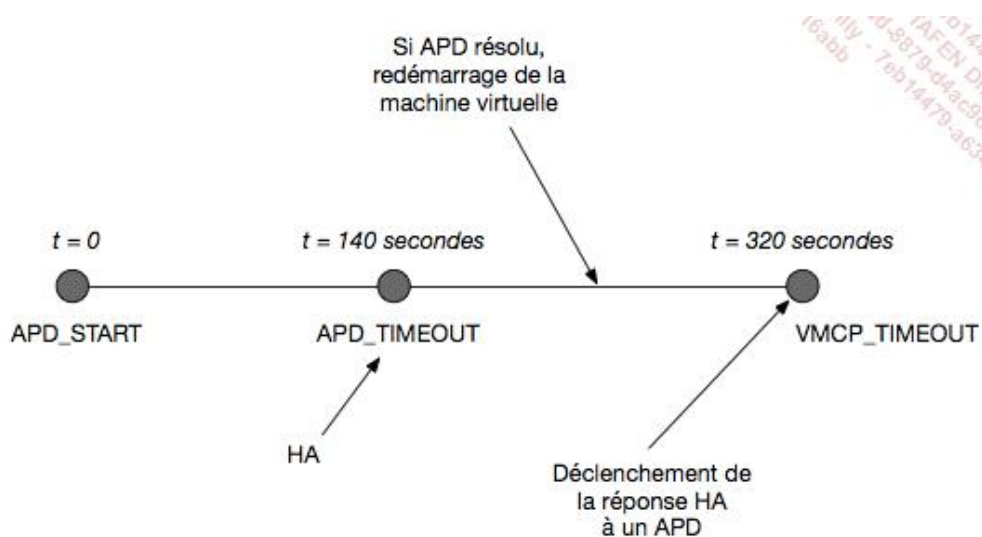
- **Le PDL ou Permanent Device Loss** qui est une perte d'accès dite non réparable. Elle se produit lorsque la baie de stockage remonte qu'un datastore n'est plus accessible pour un serveur vSphere. Cet état ne peut pas être corrigé sans l'arrêt des machines virtuelles.

Pour un événement de type PDL, les réponses paramétrables par l'administrateur sont décrites à l'aide de la capture ci-dessous :



- **Désactivé** : aucune action ne sera entreprise sur les machines virtuelles concernées.
- **Événements d'émission** : aucune action ne sera entreprise sur les machines virtuelles concernées mais des alertes seront générées pour prévenir l'administrateur.
- **Mettre hors tension et redémarrer les machines virtuelles** : l'ensemble des machines virtuelles concernées sont arrêtées. vSphere HA tentera alors de redémarrer les machines virtuelles sur des hôtes qui ont accès à la banque de données.
- **L'APD ou All Paths Down** qui est une perte d'accès dite réparable. Elle se produit lorsque tous les chemins d'accès vers un stockage sont indisponibles.

Le schéma ci-dessous présente la réaction de vSphere HA en cas d'APD :

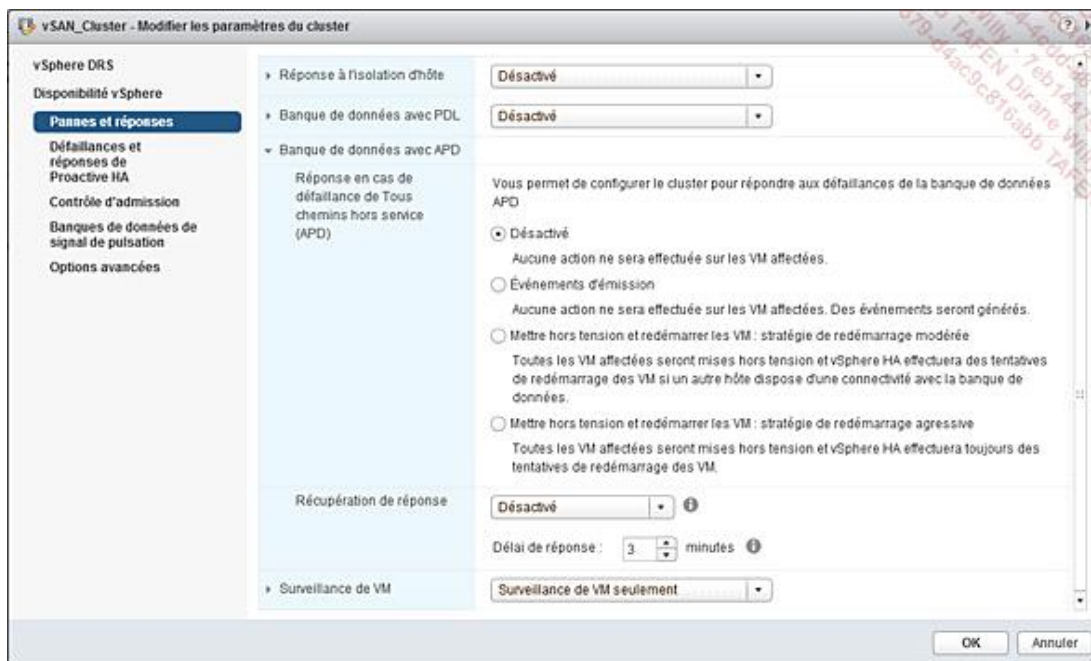


- A T0, un problème au niveau du stockage est détecté, HA lance le processus de réparation.
- Pour un événement de type APD, HA se déclenche après 140 secondes (APD_Timeout). HA commence à éliminer les

I/O des machines virtuelles présentes sur le stockage incriminé.

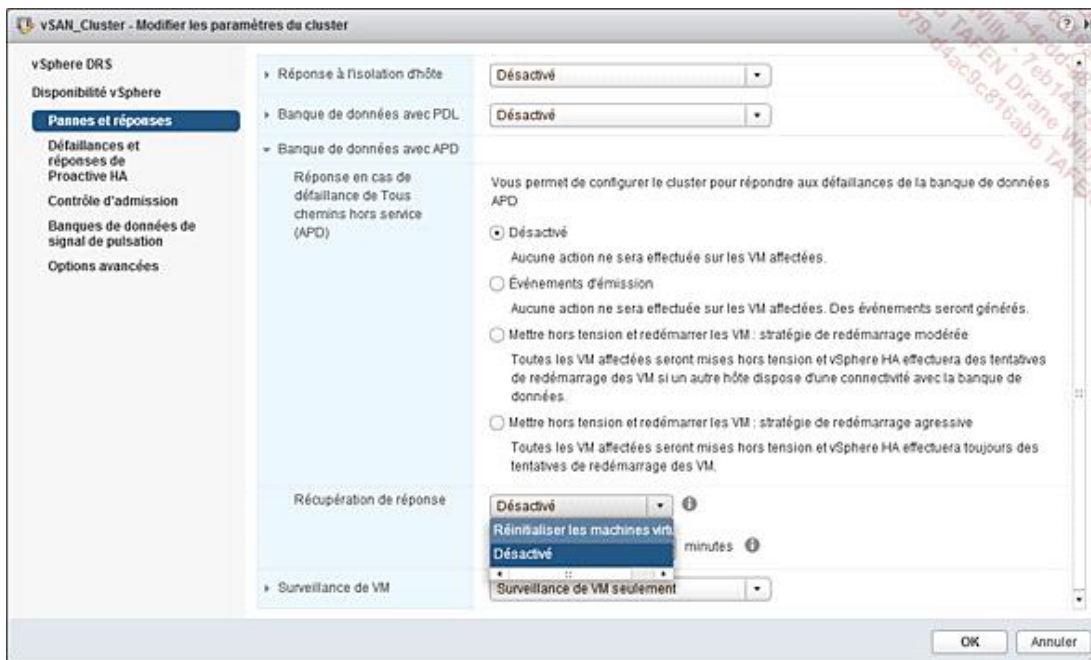
- Entre 140 et 320 secondes par défaut, si l'APD est résolu, la réponse de récupération est déclenchée. Les machines peuvent être redémarrées ou non, selon la configuration de l'administrateur dans la section « Récupération de réponse ».
- Après 320 secondes par défaut, HA prend la main et exécute la réponse à l'APD configurée par l'administrateur.

Pour un événement de type APD, les réponses (après 320 secondes) paramétrables par l'administrateur sont décrites à l'aide de la capture ci-dessous :



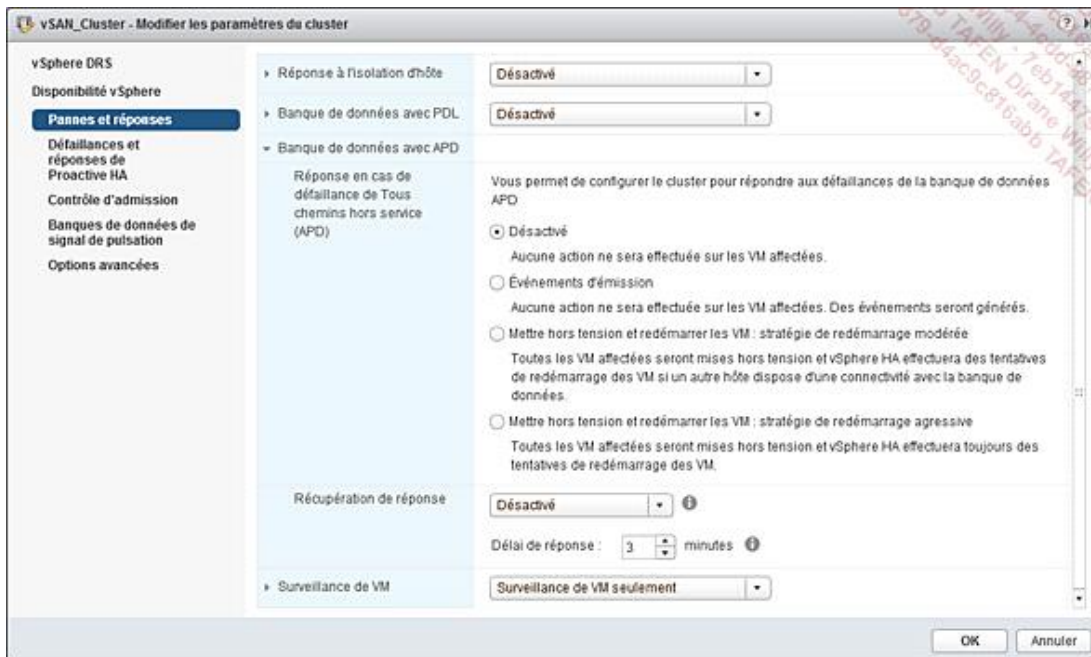
- **Désactivé** : aucune action ne sera entreprise sur les machines virtuelles concernées.
- **Événements d'émission** : aucune action ne sera entreprise sur les machines virtuelles concernées mais des alertes seront générées pour prévenir l'administrateur.
- **Mettre hors tension et redémarrer les machines virtuelles - stratégie modérée** : l'ensemble des machines virtuelles concernées seront arrêtées et vSphere HA tentera alors de redémarrer les machines virtuelles sur d'autres hôtes **si et seulement si** ces autres hôtes ont accès à la banque de données. Cela évite d'éteindre la machine virtuelle si aucun autre hôte ne peut accéder à la banque de données.
- **Mettre hors tension et redémarrer les machines virtuelles - stratégie agressive** : l'ensemble des machines virtuelles concernées seront arrêtées et vSphere HA tentera **toujours** de redémarrer les machines virtuelles, indépendamment de l'accès d'autres hôtes à la banque de données, ce qui peut entraîner un non-redémarrage de la machine virtuelle si aucun autre hôte n'a accès à la banque de données.

Enfin, l'administrateur peut configurer la réponse de récupération si l'APD est résolu entre 140 et 320 secondes par défaut à l'aide de la dernière section. Comme vous le voyez, l'administrateur peut décider selon ce qui s'affiche :



- **Désactivé** : aucune action ne sera entreprise sur les machines virtuelles concernées.
- **Réinitialiser les machines virtuelles** : les machines virtuelles seront redémarrées.

Enfin, il est à noter que la dernière section nous permet de définir le temps de réponse entre l'attente de la résolution de l'APD et la réponse à l'APD. Par défaut, comme énoncé plus haut, vous voyez que la valeur est de 3 minutes, c'est-à-dire 180 secondes.



h. Contrôle d'admission et stratégies associées

Le serveur vCenter utilise les stratégies de contrôle d'admission afin que le cluster puisse garder suffisamment de ressources dans le cadre d'une bascule automatique et de vérifier que les ressources réservées pour les machines virtuelles sont bien garanties.

Le contrôle d'admission peut intervenir dans les scénarios suivants :

- Mise sous tension d'une machine virtuelle
- Migration d'une machine virtuelle
- Augmentation de la réservation processeur / mémoire d'une machine virtuelle

La capacité en ressource du contrôle d'admission se base sur les ressources disponibles au niveau du cluster. Cela implique les choses suivantes :

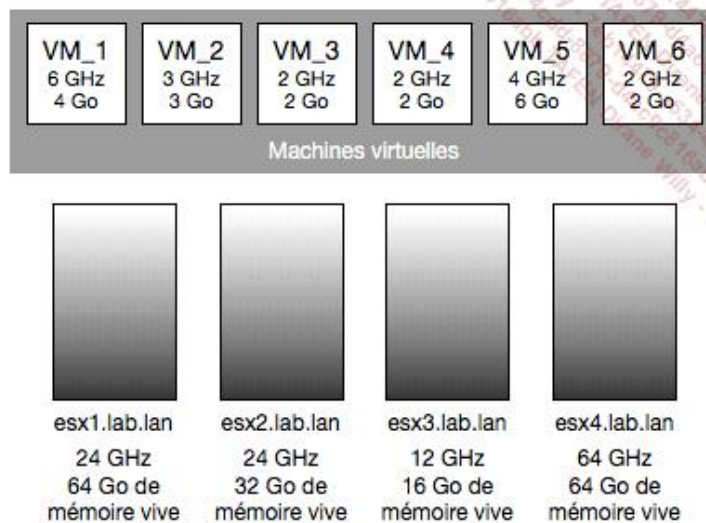
- Si un serveur est indisponible, il reste comptabilisé dans les ressources du cluster.
- Si un serveur est en maintenance ou en standby, il n'est plus comptabilisé dans les ressources du cluster.

Le contrôle d'admission s'effectue à trois niveaux :

- **Du serveur**, afin qu'il y ait suffisamment de ressources disponibles pour les machines virtuelles exécutées sur ledit serveur.
- **Du pool de ressources**, afin qu'il y ait suffisamment de ressources disponibles afin de satisfaire les réservations, parts et limites de toutes les machines virtuelles associées à celui-ci.
- **De HA**, qui s'assure qu'il y ait suffisamment de ressources réservées dans le cluster afin d'autoriser le redémarrage des machines virtuelles en cas de perte de serveurs vSphere (failover capacity).

Au niveau du cluster on parle de **stratégies de contrôle d'admission**, qui sont au nombre de trois :

Tout d'abord, le premier cas où la stratégie de contrôle d'admission est basée **sur une politique de slots**. Prenons un schéma pour expliquer son fonctionnement.



Ce schéma décrit une architecture dotée de 4 hôtes et de machines virtuelles s'exécutant au sein du cluster.

Dans ce scénario, la stratégie de contrôle d'admission est exécutée en plusieurs étapes :

- La taille d'un slot est calculée en fonction de la consommation processeur et mémoire maximale à travers l'ensemble du cluster. Ici, notre taille de slot sera fixée à **6 GHz / 6 Go**.

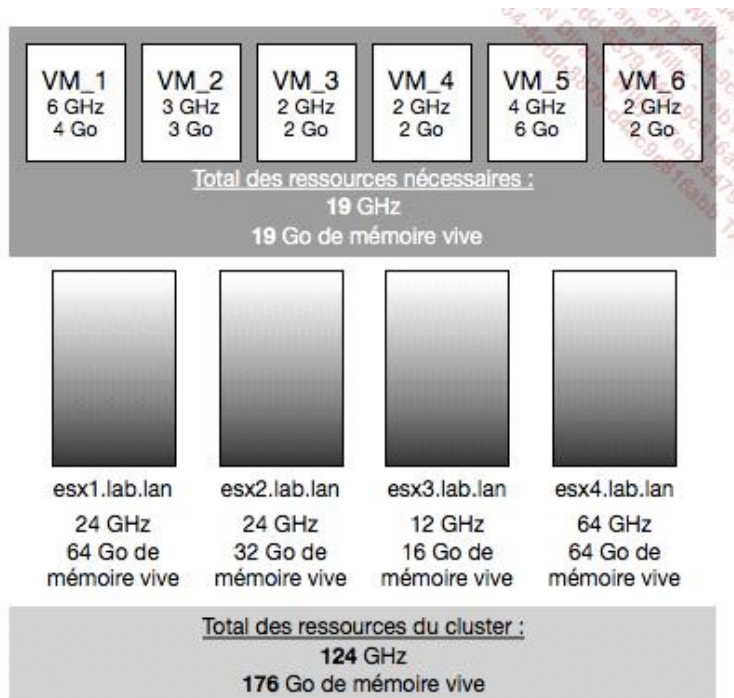
➤ La taille des slots peut aussi être fixée, ou encore évaluée en prenant en compte les réservations. Attention dans ce cas précis : le slot concerne la mémoire ET le processeur. Pour obtenir la taille du slot, vCenter prendra en référence la plus grosse réservation sur une VM pour la mémoire vive, et de même pour le processeur. Vous pouvez donc vous retrouver avec une évaluation des capacités de failover du cluster complètement farfelue à cause de quelques réservations configurées sur les VMs. Il est donc conseillé de configurer les réservations sur les resource pools plutôt que sur les VMs directement.

- On calcule ensuite le nombre de slots que peuvent supporter les hôtes en fonction de la taille de slot déterminée. Ici, voici le nombre de slots supportés par hôtes :
 - esx1.lab.lan : 4 slots
 - esx2.lab.lan : 4 slots
 - esx3.lab.lan : 2 slots
 - esx4.lab.lan : 10 slots
- Au total, le cluster dispose donc de **20 slots** dont 14 de disponibles.
- Si un hôte tombe en panne, ce nombre sera nécessairement réduit. En fonction de l'hôte qui dysfonctionne, la capacité d'admission sera plus ou moins importante. Si le plus important hyperviseur de notre cluster (esx4.lab.lan) tombe, il ne nous restera que 10 slots au total dans notre cluster, dont quatre de disponibles.

Souvenez-vous bien de cela : pour calculer la capacité de failover, HA calcule toujours « au pire des cas » donc si le ou les serveurs les plus puissants tombent. Ceci explique que bien souvent on se retrouve avec une capacité de failover bien moindre qu'anticipée.

➤ Le fonctionnement de HA nous pousse à appliquer une bonne pratique : uniformiser les capacités matérielles des serveurs faisant partie du même cluster. Idéalement on utilise des serveurs rigoureusement identiques.

Intéressons-nous maintenant à la stratégie de contrôle d'admission basée sur **le pourcentage de ressources réservées**. Reprenons un schéma pour en expliquer son fonctionnement :



Le schéma décrit à nouveau une architecture à 4 hôtes avec des machines virtuelles dotées de réservations processeur et mémoire.

Dans ce scénario, le respect de la stratégie de contrôle d'admission est assuré de la façon suivante :

- Le premier calcul relatif tout d'abord aux ressources dont ont besoin les machines virtuelles allumées dans le cluster. Pour ce calcul, vSphere HA utilise dans ce cas les réservations configurées sur les machines virtuelles. Concernant le calcul mémoire, il sera ajouté la surcharge mémoire (*memory overhead*), nécessaire pour exécuter les machines dans notre cluster. Si les machines virtuelles sont configurées sans réservation, des valeurs par défaut de 0 Mo et 32 MHz sont appliquées. Ces paramètres peuvent être changés à l'aide des options de configuration avancées (https://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033250) du contrôle d'admission *das.vmmemoryminmb* et *das.vmcputminmhz*.
- On calcule ensuite la capacité totale disponible à travers le cluster.
- Le calcul suivant consiste à déterminer la capacité de *failover* concernant les ressources de type processeur et mémoire.
- Enfin, on détermine si la capacité de *failover* processeur et mémoire est inférieure au pourcentage configuré par l'administrateur si l'action est validée. Si c'est le cas, le contrôle d'admission refusera l'action entreprise.

Dans notre cas, le calcul de la capacité de failover processeur est le suivant :

$$\frac{124 \text{ GHz} - 19 \text{ GHz}}{124 \text{ GHz}} = 84.6 \%$$

Le calcul de la capacité de *failover* mémoire est le suivant :

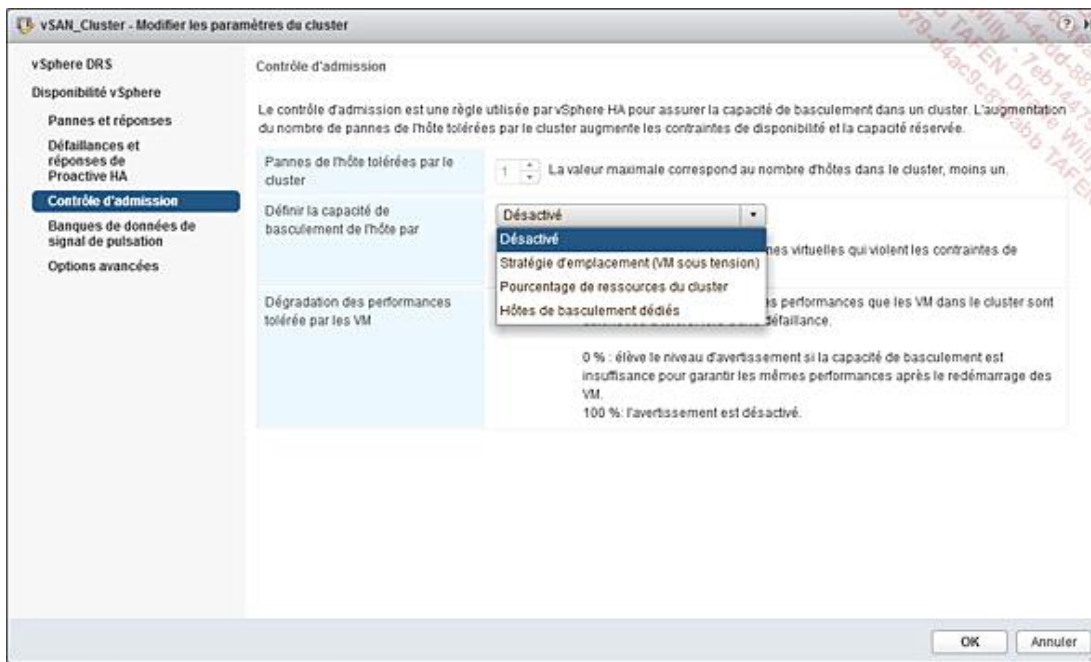
$$\frac{176 \text{ Go} - 19 \text{ Go}}{176 \text{ Go}} = 89.2 \%$$

Si l'administrateur avait fixé une capacité de *failover* de 40 % pour les ressources processeur et mémoire, le cluster serait encore en capacité d'admettre des machines virtuelles, tant que ce seuil n'est pas en dessous de sa valeur pour les deux types de ressources.

En cas de perte de serveurs, ce pourcentage serait nécessairement amené à changer, en fonction des hôtes potentiellement amenés à tomber en panne.

Enfin, la troisième stratégie de contrôle d'admission consiste à **dédier des hôtes** pour ne servir qu'en cas de panne de serveurs (*failover hosts*). Vous ne pourrez pas exécuter de machines virtuelles sur ces hôtes dans des conditions de fonctionnement nominales. De plus si DRS est utilisé, le serveur de réserve (failover host) ne sera pas utilisé dans le calcul des ressources disponibles, et il n'y aura pas de migration vMotion vers ce serveur.

Ces stratégies de contrôle d'admission ainsi que le nombre de pannes d'hôte à tolérer dans le cluster peuvent être configurés à l'aide de l'écran suivant :



Le contrôle d'admission impose des contraintes sur l'usage des ressources et aucune violation de ces contraintes n'est permise. L'administrateur peut décider de désactiver le contrôle d'admission en sélectionnant l'option « Désactivé » sur l'écran ci-dessus.

Dans cette configuration, l'administrateur ne dispose d'aucune garantie que l'ensemble des machines virtuelles pourront être redémarrées en cas de perte d'un serveur vSphere.

Voici les principales raisons pour violer les contraintes de la stratégie du contrôle d'admission :

- Les mises à jour via VUM car elles peuvent entraîner la mise en mode de maintenance d'un ou plusieurs hôtes et donc entraîner une indisponibilité temporaire de certains serveurs qui entraîne à son tour une modification des ressources dont dispose le cluster.
- Les opérations de maintenance

i. Groupe de machines virtuelles et priorité de redémarrage

Depuis vSphere 6.5, la fonction Orchestrated restart est proposée.

Quand bien même vSphere HA est activé, l'administrateur est libre de définir des groupes de machines virtuelles (qu'on peut aussi utiliser avec DRS) ainsi que des règles pour forcer les comportements suivants :

- Garder les machines ensemble
- Séparer les machines virtuelles
- Machines virtuelles sur les hôtes
- Machines virtuelles à machines virtuelles

Il faut cependant prêter attention à ce que ces règles n'entravent pas la disponibilité des machines virtuelles, en empêchant notamment vSphere HA d'agir.

Par la dernière option, un administrateur peut influencer l'ordre de redémarrage des machines virtuelles, comme l'illustre la capture suivante.

vSAN_Cluster - Créer une règle de VM/hôte

Nom :

☒ Activer la règle.

Type :

Description :

Les machines virtuelles du groupe de VM DHCP_SERVERS seront redémarrées en premier. Les machines virtuelles du groupe de VM LINUX_SERVERS seront redémarrées ensuite, une fois la condition de redémarrage de dépendance de cluster remplie.

Commencez par redémarrer les VM du groupe de VM :

Redémarrez ensuite les VM du groupe de VM :

OK Annuler

L'administrateur souhaite ici démarrer au préalable les serveurs DHCP avant de démarrer les serveurs de production Linux.

La capture suivante vous montre des règles en production.

HomeLabCluster

Démarrage Résumé Surveiller **Configurer** Autorisations Hôtes VM Banques de données Réseaux Update Manager

Services

- vSphere DRS
- Disponibilité vSphere
- Virtual SAN
- Général
- Gestion de disques
- Domaines de pannes et cluster étendu
- Santé et performances
- Cibles iSCSI
- Groupes d'initiateurs iSCSI
- Aide à la configuration
- Mises à jour
- Configuration
- Général
- Attribution de licence
- VMware EVC
- Groupes de VM/hôte
- Règles de VM/hôte**

Règles de VM/hôte

Ajouter... Modifier... Supprimer

Nom	Type	Activée	Conflits	Définie ...
RULE_PROD_1	Exécuter des VM sur la base de la...	Oui	0	Util
RULE_PROD_2	Exécuter des VM sur la base de la...	Oui	0	Util
RULE_PROD_3	Exécuter des VM sur la base de la...	Oui	0	Util

Détails de la règle de VM/hôte

vSphere HA commencera par redémarrer les machines virtuelles du groupe de VM MAIL_INFRA. Une fois les conditions de redémarrage de dépendance de cluster remplies, les machines virtuelles du groupe de VM FRONTEND_VM seront lancées.

Ajouter... Supprimer

MAIL_INFRA Membres du groupe

- gra1_Exchange2K13
- home_SpamAssassin
- gra1_MXFailover

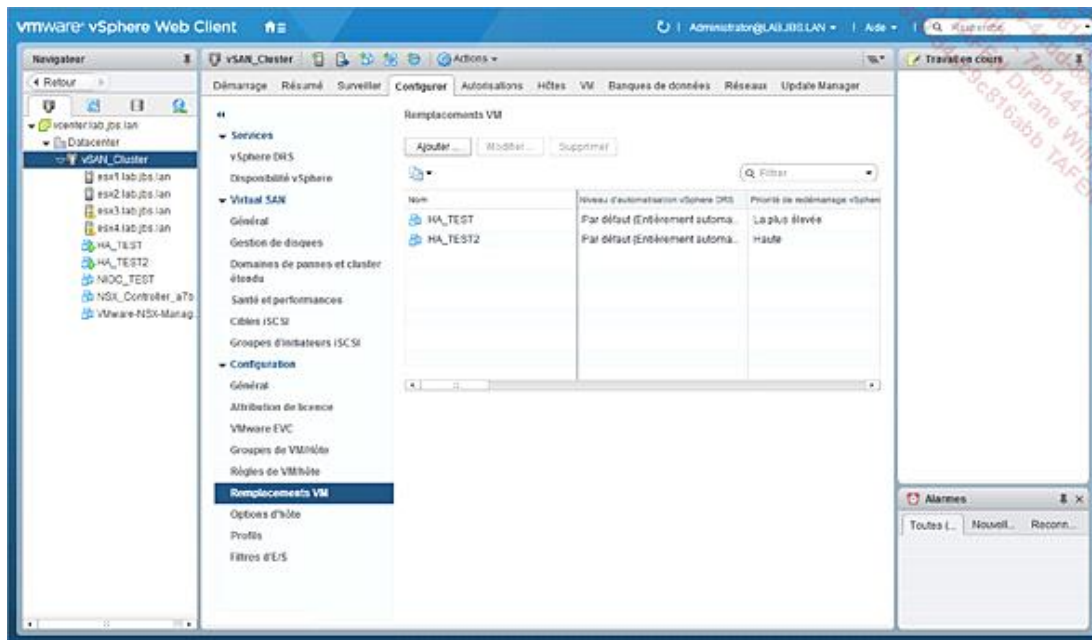
Ajouter... Supprimer

FRONTEND_VM Membres du groupe

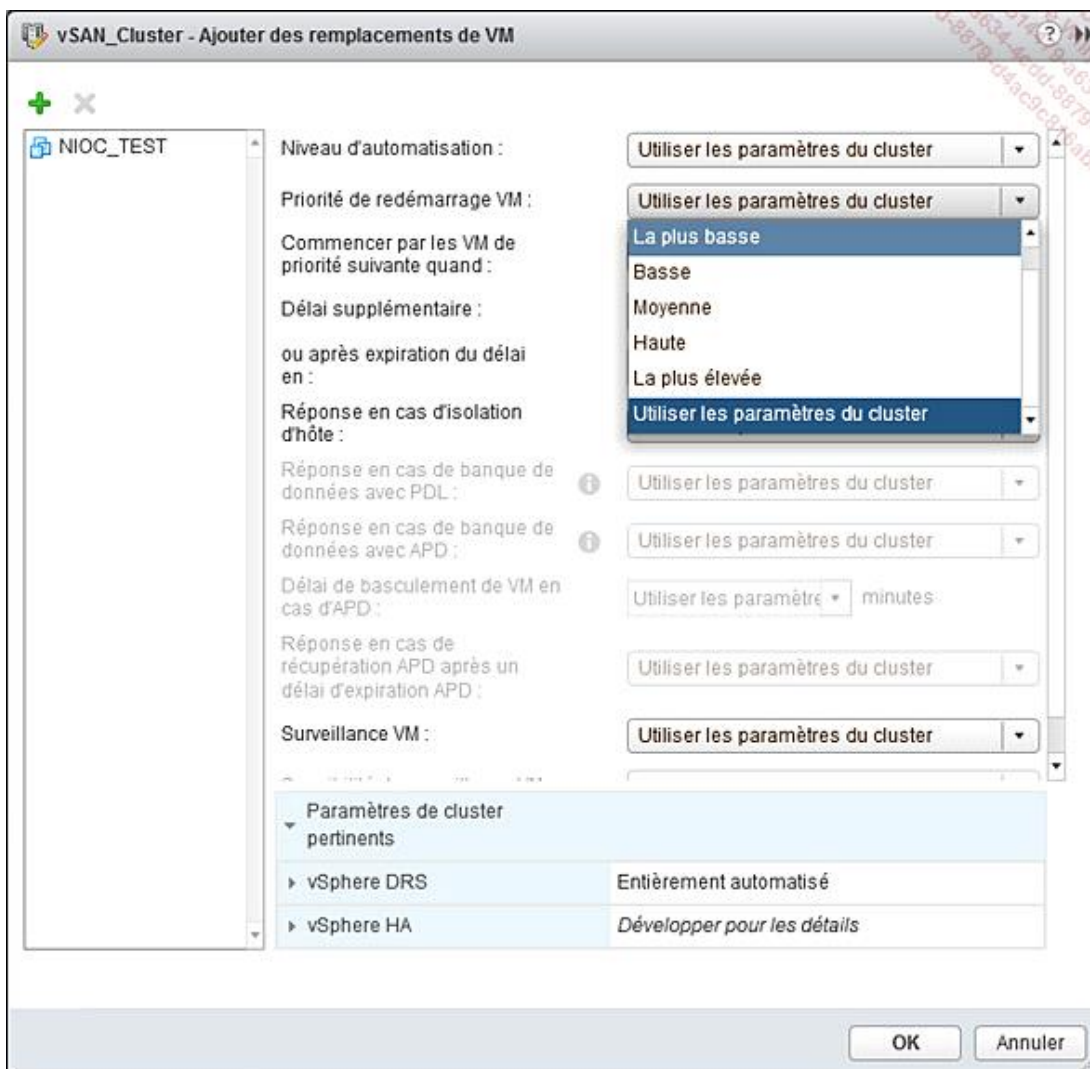
- home_Nginx

Une autre façon d'influencer l'ordre de redémarrage de machines virtuelles est de définir la priorité de redémarrage.

Il reste possible d'utiliser les priorités directement sur les VM, comme dans les versions antérieures de vSphere. À l'aide de la section **Remplacement VM**, l'administrateur peut définir des règles de priorité directement sur les machines virtuelles.



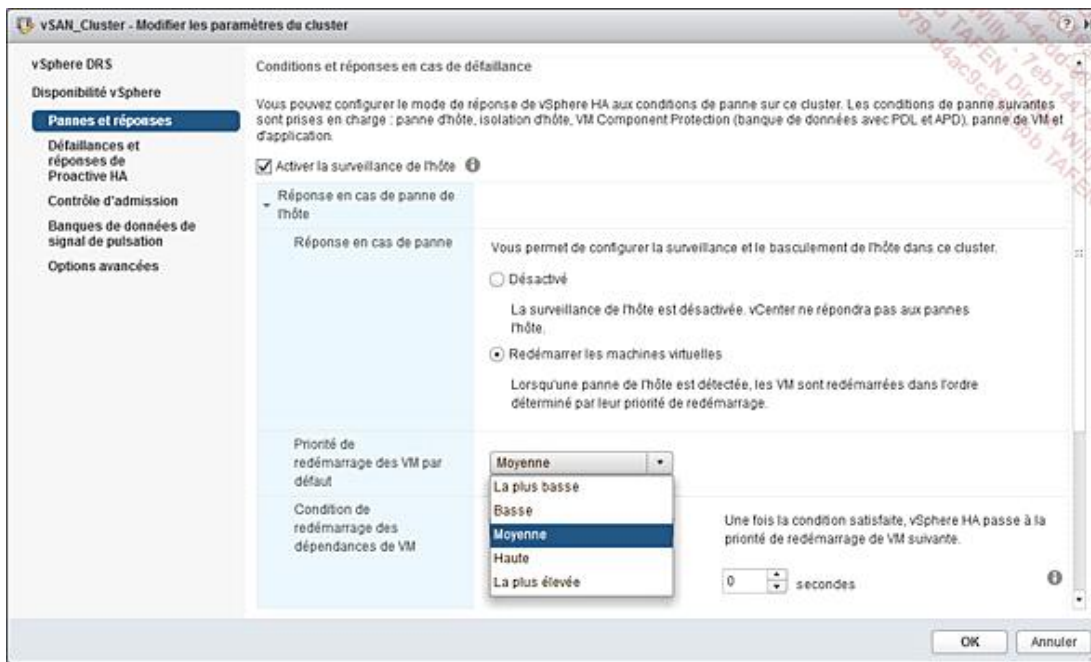
Vous pouvez définir la priorité des machines virtuelles parmi : **La plus élevée, Haute, Moyenne** (la valeur par défaut dans le cluster), **Basse, La plus basse**, comme vous pouvez le voir dans la capture suivante.



L'administrateur peut d'ailleurs sur cet écran définir le critère qui déterminera quand le redémarrage de la prochaine machine aura lieu, grâce à l'option **Commencer par les VM de priorité suivante quand**. L'administrateur est libre de choisir entre autres, entre l'allocation des ressources ou la détection de heartbeats en provenant des VMware Tools ou de heartbeats applicatifs.

Comme attendu, les machines virtuelles dotées d'une priorité plus haute seront redémarrées en priorité par rapport aux autres.

La priorité par défaut qui s'applique au niveau du cluster peut être définie dans la configuration du cluster, comme l'illustre la capture ci-dessous.



Sur ce même écran, l'administrateur peut également spécifier quand la prochaine machine sera redémarrée.

j. Perte de performance acceptable

Il existe un autre seuil que l'administrateur peut définir pour avoir une maîtrise plus fine de son infrastructure en cas de panne, celui de la dégradation des performances acceptables par les machines virtuelles. L'objectif est de pouvoir obtenir une alerte si la dégradation des performances est supérieure au seuil fixé.

Le calcul est mené à la fois concernant les ressources processeur et mémoire et le seuil déterminé par l'administrateur. Les ressources mémoires sont considérées de la perspective de la mémoire réservée (et non celle réellement consommée) et la surcharge mémoire. Par exemple, un seuil fixé à 50 % entraînera le calcul suivant :

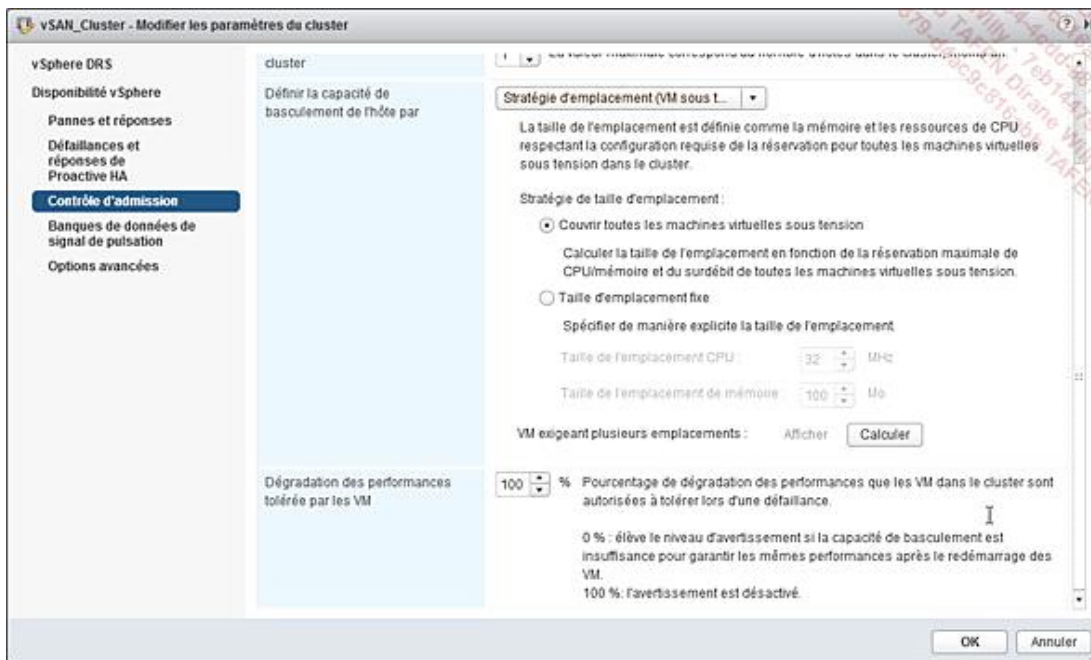
$$\text{baisse de performance acceptable} = \text{utilisation courante} * 50\%$$

Si l'utilisation courante moins la baisse de performance acceptable excède la capacité disponible, une alerte est générée.

Plusieurs valeurs types peuvent être définies :

- 100 % (valeur par défaut) - Aucune alerte n'est ou ne sera produite.
- 0 % - un avertissement est généré aussi tôt que l'utilisation du cluster excède la capacité disponible
- Une valeur définie par l'administrateur

Cette valeur peut être définie à l'aide de l'écran suivant.



Il est à noter que ce seuil d'alerte n'est disponible que si DRS est activé au sein du cluster.

k. HA et les autres composants vSphere

vSphere HA peut être utilisé en combinaison avec d'autres technologies VMware, comme nous allons le voir.

I. DRS

DRS peut, comme cela est fait dans un fonctionnement nominal, participer au rééquilibrage du cluster après la panne d'un hôte. Cela peut passer par l'allumage d'autres hôtes mis en veille jusqu'ici par le biais de DPM.

Selon les cas, il se peut que vSphere HA ne soit pas capable de garantir le redémarrage des machines virtuelles. La liste non exhaustive de ces scénarios est présentée ci-dessous :

- Si DPM est activé mais que le contrôle d'admission est désactivé. Dans ce cas, il se peut que les machines virtuelles soient consolidées sur un nombre d'hôtes réduit, les autres hôtes étant placés en veille et réduisant de fait la capacité d'accueil du cluster en cas de panne.
- Si l'administrateur a configuré des règles d'affinité « VM-Host » qui peuvent limiter de fait les hôtes vers lesquels les machines peuvent s'exécuter.
- La capacité du cluster agrégée peut être suffisante, mais si elle est fragmentée entre plusieurs hôtes, il se peut que vSphere HA ne puisse pas redémarrer les machines virtuelles. Par exemple, s'il y a 10 Go de mémoire vive disponible sur le cluster, mais répartis sur deux serveurs hôtes, une machine virtuelle ayant une réservation de mémoire vive fixée à 10 Go ne pourra pas démarrer dans le cas d'un contrôle d'admission strict.

Dans ces cas, DRS peut jouer un rôle dans une réaction à la panne adaptée, pour permettre à HA de fonctionner correctement. Une évaluation peut être déclenchée (RUN DRS) afin d'éventuellement migrer certaines machines virtuelles pour libérer assez de ressources sur un serveur (ou un nombre restreint de serveurs).

m. vSAN

Vous pouvez utiliser conjointement vSphere HA avec vSAN, à la condition que vous disposiez d'au moins trois hôtes dans un cluster HA, exécutant au minimum vSphere 5.5.

Cependant, il faut noter plusieurs différences notables qui influencent le fonctionnement de vSphere HA quand il est utilisé avec vSAN.

La première concerne le réseau utilisé par HA. vSAN utilise son propre réseau pour fonctionner par le biais d'interfaces VMkernel le plus souvent dédiées. Dans le cas où vSphere HA serait activé dans un cluster vSAN existant, le réseau utilisé par vSphere HA sera le même que celui utilisé par vSAN, et non plus le réseau de gestion traditionnel.

L'ensemble des communications effectuées dans le cadre de vSphere HA utiliseront donc cette interface. Cela implique qu'en cas de panne et notamment d'un isolement d'hôte, la vérification sera effectuée sur l'accessibilité du réseau vSAN au lieu du réseau de gestion de base.

Une des autres différences concerne les banques de données utilisables pour tester la connectivité des hôtes à ces dernières. Il est à noter qu'une banque de données vSAN ne peut pas être utilisée pour les *heartbeats*.

Depuis le début de ce chapitre, nous avons abordé les différentes caractéristiques de vSphere HA et son fonctionnement. Mais existe-t-il un moyen encore plus efficace de protéger les machines virtuelles ? C'est ce que nous allons découvrir maintenant !

2. Fault Tolerance

vSphere HA propose un moyen fiable de garantir la haute disponibilité d'une machine virtuelle.

Il ne vous a cependant pas échappé qu'en cas de panne d'un hôte, la machine virtuelle est arrêtée **brutalement** pour être ensuite **redémarrée** sur un autre hôte du cluster, ce qui peut entraîner des pertes de données mais aussi un temps d'indisponibilité certain.

Essayez d'imaginer dans votre cluster une machine virtuelle exécutant un service de base de données (MySQL, MSSQL...) **critique**. Cette dernière subit une panne d'hôte en pleine écriture. Quand bien même il existe des mécanismes applicatifs pour garantir l'intégrité des données, l'administrateur n'aura aucune garantie que la base de données pourra fonctionner en redémarrant la machine virtuelle sur un autre hôte.

Visualisez maintenant un service s'exécutant sur une machine virtuelle qui met une heure à démarrer à froid, cela sera autant de temps d'indisponibilité à subir pour les utilisateurs de ce service.

Enfin, positionnez-vous comme un hébergeur qui garantit une disponibilité à 5 chiffres, c'est-à-dire 99,999 % du temps. Pour rappel, cette valeur autorise à l'hébergeur un temps d'indisponibilité annuel de... 5.26 minutes. Au-delà, le contrat de service (SLA - *Service Level Agreement*) mentionne bien souvent l'indemnisation du client de la part de l'hébergeur si jamais ce contrat n'est pas respecté.

Vous le voyez, il y a des scénarios où même l'administrateur ne peut se permettre un arrêt brutal d'une machine virtuelle. Pour répondre à ce besoin, VMware propose une solution : Fault Tolerance (appelée également FT ou tolérance à la panne).

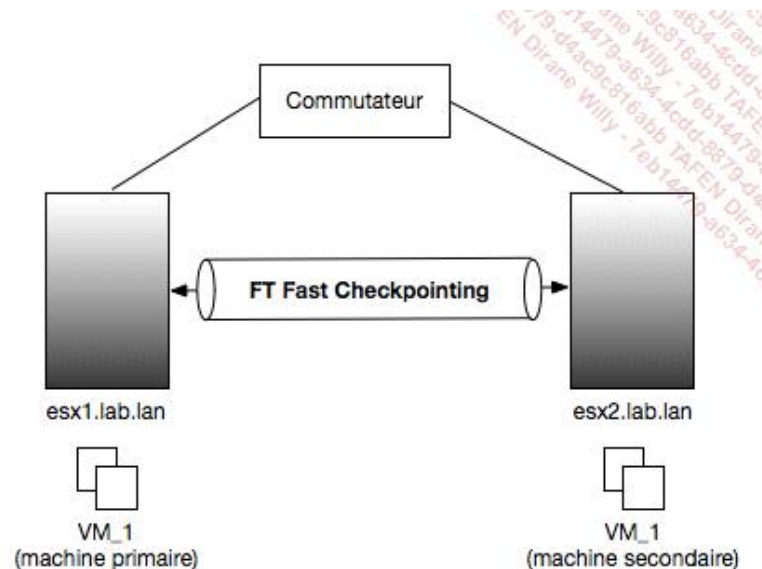
Grâce à cette technologie, l'administrateur peut garantir une disponibilité **continue** de la ressource. En cas de panne, la machine virtuelle continue à s'exécuter et n'est pas arrêtée. Aucune donnée, connexion réseau ou transaction n'est perdue.

Comment cela fonctionne ? C'est simple ! Il s'agit d'exécuter la machine virtuelle ... deux fois ! La machine est donc exécutée simultanément sur deux hôtes. Les deux instances portent le nom de machine virtuelle primaire et secondaire.

L'hôte exécutant la machine virtuelle primaire envoie en permanence des informations représentant les changements opérés au sein de la machine primaire à destination de l'hôte exécutant la machine secondaire.

L'objectif est de permettre une synchronisation totale d'état entre la machine primaire et secondaire.

Le schéma explique comment fonctionne Fault Tolerance pour une machine virtuelle s'exécutant sur un hôte :



Le cluster peut bénéficier de cette technologie si les hôtes utilisent au minimum un processeur Intel Sandy Bridge, AMD Bulldozer ou plus récent.

VMware recommande de dédier une connexion 10 Gbit/s entre les deux hôtes supportant la machine primaire et secondaire et d'en modifier la MTU pour permettre les jumbo frames. Sur cette connexion, il est recommandé de disposer d'une latence basse. Les connexions dédiées à Fault Tolerance peuvent être assurées par le biais d'un commutateur ou à l'aide de connexions directes entre les hôtes.

Il est à noter que depuis vSphere 6.0, la technologie qui fait fonctionner Fault Tolerance utilise le Fast Checkpointing, comparable avec les mécanismes de vMotion qui synchronisent l'état d'exécution, la mémoire, le stockage et le réseau entre les deux instances primaires et secondaires. Auparavant, il était fait appel à un mécanisme d'enregistrement s'opérant sur la machine primaire et de rejeu sur la machine secondaire (vLockstep).



On peut considérer la fonctionnalité de FT (legacy) comme une migration vMotion qui ne se termine pas.

L'une des autres nouveautés de vSphere 6.0 est le SMP-FT (*Symmetric Multi-Processing Fault Tolerance*). SMP-FT vous permet de garantir une disponibilité continue d'une machine virtuelle dotée de 4 vCPU et de 64 Go de mémoire vive quand l'ancien mode (Legacy FT) ne supportait qu'un vCPU.



On peut considérer la fonctionnalité de SMP-FT comme une migration vMotion shared-nothing qui ne se termine pas.

Enfin, sachez que vSphere 6.5 fixe à 98 le nombre maximal de machines virtuelles pouvant bénéficier de Fault Tolerance au sein d'un cluster. Sur l'ensemble de ces machines virtuelles, 256 vCPUs maximum sont supportés dans le cluster.

Passons désormais à la pratique !