Cluster VMware ESX 4.1

(22 mars 2011)

1 Objet

Mise à disposition de 5 serveurs Windows 2008, version Datacenter 64 bits, de façon hautement disponible. Et ce à l'aide de la virtualisation VMware : serveurs virtualisés (VM = Virtual Machine).

2 Machines physiques utilisées

Deux lames, modèle Power Edge M610, bios 2.0.14.
Chacune de ces lames est composée de :

Processeurs

Modèle : Intel(R) Xeon(R) CPU E5530 @ 2.40GHz
Vitesse du processeur : 2,4 GHz
Sockets du processeur : 2
Coeurs de processeur par socket : 4
Processeurs logiques : 16
Hyperthreading : Activé

RAM

32 GO (par lames).

Ethernet

4 cartes : 1 Gbit, Broadcom NetXtreme II BCM5709.

Stockage

Raid 1 (80 GO) sur chaque lame (Dell SAS 6/iR Integrated).
LPe 12000 8 Gbit Fibre channel vers baie SAN, via 4 chemins redondants.
Mis à disposition depuis la baie SAN : un raid 5 sur 7 disques de 450 GO chacun, sur lequel on créé des LUNs de tailles choisies pour chacune des VMs.

3 Installation

Le cluster de virtualisation VMware est composé de deux socles ESX 4.1, qui sont installés chacun sur une des deux lames.
Ce cluster est géré par un serveur VCenter VMware (4.1), installé sur une VM avec un OS Windows 2008 Datacenter 64 bits. Pour l'instant un backup (dormant) de ce serveur VCenter réside sur une

lame physique sous Windows 2008. La question du choix de l'emplacement du VCenter (machine physique ou VM) est discutée ci-dessous paragraphe 5.
Les licences VMware ESX et VCenter sont installées, leur validité est permanente (au moins pour cette version de vSphere VMware : la 4.1, dernière en date).

Les 5 machines virtuelles (VMs) créées dans le cluster sont des Windows 2008 Datacenter 64 bits. Leurs licences doivent être activées sous 60 jours, période renouvelable 3 fois. Soit au total 240 jours possibles sans activation.

Au 21/03, il reste donc : 35 + 180 jours avant l'activation.

En plus du serveur VCenter virtualisé, les 5 VMs sont installées comme suit :

- Win2008-Tomcat : 192.168.4.129, RAM : 2 GO, disques C : 60 GO ; E : 10 GO
- Win2008-Oracle : 192.168.4.130, RAM : 2 GO, disques C : 60 GO ; E : 100 GO
- Win2008-Carto-prod : 192.168.4.131, RAM : 5 GO, disques C : 60 GO ; E : 20 GO
- Win2008-Apache : 192.168.4.132, RAM : 2 GO, disques C : 60 GO ; E : 10 GO
- Win2008-Carto-demo : 192.168.4.139, RAM : 5 GO, disques C : 60 G0 ; E : 20 GO.

Elles sont accessibles en TSE (user : Administrateur, passwd : Eliot_38). En parallèle les comptes Raphael et Florian ont été créés pour ne pas perturber une connexion Administrateur active.

Est installé sur chacune de ces VMs, les "VMware Tools", paquet logiciel quasiment indispensable à ne pas désinstaller :

1) ils fournissent des drivers Windows pour le hardware simulé : carte réseau virtuelle (vmxnet3), disque LSI logic SAS, horloge système, carte video (pour la console), souris...
2) ils permettent d'optimiser la vitesse de reboot des VMs, (lorqu'un reboot de la VM est demandé),
3) ils permettent la synchronisation des horloges des VMs sur celles des socles ESX. Ces dernières sont réglées sur un serveur de temps (NTP) du LAN Eliot local.
4) ils permettent la surveillance par le cluster ESX des VMs : soit en cas de défaillance de l'OS (écran bleu : "BSOD"), soit en cas de système inutilisable pour cause de saturation de l'utilisation des ressources (en RAM principalement : "fork bomb" par exemple). Dans l'un de ces deux cas, les pulsations émises par les VMware Tools ne sont plus reçues par le socle ESX hébergeant la VM défaillante, et l'on peut choisir de rebooter la VM (selon configuration au niveau du cluster).

4 Problématique RAM et CPU

D'abord un retour sur les deux différents modes de disponibilité proposés par VMware cluster.

Mode 1 : le mode par défaut ("HA" : High Availability).

Dans ce mode, la disponibilité des VMs est "accrue" mais pas totale. On installe les VMs en équilibrant la charge dans le cluster (par exemple 2 VMs sur le socle 1, et 2 VMs sur le socle 2). En cas de panne de l'un des socles, toutes les VMs de ce socle sont basculées automatiquement sur l'autre socle (ou sur les autres socles si on en a plus de deux).
Cela implique un reboot (optimisé) des VMs qui ont été basculées, et le temps de coupure du service pour ces VMs (pour un Windows 2008 server vierge) est de moins d'une minute.

Mode 2 : le mode "protégé" ("HA" + "FT" avec VLockstep, FT = Fault Tolerance).

Ici, en cas de panne de l'un des deux socles, il n'y a pas de reboot des VMs, la continuité du service est totale. Il y a juste un ralentissement de moins d'une seconde, et au niveau réseau la perte d'un ping (mais pas la perte de connexions réseau actives).
Ce mode impose des conditions de ressources physiques très fortes, en RAM, en CPU, et aussi au niveau réseau avec une interface par lame dédiée (voire plus si on a beaucoup de VMs à protéger).
La liste des conditions à remplir se trouve par exemple ici :
http ://communities.vmware.com/blogs/vmroyale/2009/05/18/vmware-fault-tolerance-requirements-and-limitations
Notamment ce mode n'est pas optimisé pour des applications conçues spécifiquement pour fonctionner en mode SMP, car l'on est obligé de créer les VMs avec un seul vCPU. Cela étant, ce sont les socles sous-jacents qui font du round-robin sur les CPU physiques disponibles, et on peut mettre des priorités CPU aux VMs qui en ont le plus besoin.
En termes de performances, cette question de l'unique vCPU en mode HA+FT ne semble pas problématique pour toutes les applications non spécifiquement conçues pour du SMP.

Voir par exemple :
http ://itknowledgeexchange.techtarget.com/virtualization-pro/masters-guide-to-vmware-fault-tolerance/
http ://blog.peacon.co.uk/understanding-the-vcpu/
http ://kb.vmware.com/selfservice/microsites/search.do ?language=en_US&cmd=displayKC&externalId=1010184

5 Emplacement du serveur VCenter

Selon différentes sources, dont VMware eux-mêmes, l'installation du serveur VCenter qui gère le cluster est préconisée dans une VM. Voir par exemple :

http ://www.vmwareinfo.com/2009/01/vmware-virtualcenter-as-virtual-machine.html
http ://www.vmware.com/pdf/vi3_vc_in_vm.pdf

Bien qu'a priori paradoxal, cela fonctionne très bien, l'intérêt principal étant certainement d'obtenir une haute disponibilité du VCenter.
Les principaux points d'interrogation à ce mode fonctionnement étant les suivants :

1) Si la VM hébergeant VCenter crashe, quid du cluster et des VMs ?

Cluster et VMs continuent de fonctionner normalement, même en mode HA + FT.

2) Comment rebooter la VM hébergeant VCenter dans le cas 1) ?

On se connecte directement sur l'hôte ESX qui héberge cette VM, et on la redémarre. Bien sûr si elle est surveillée, le reboot est automatique.

3) Et s'il y a un problème d'accès réseau à la VM hébergeant VCenter, sans console pour cette VM, comment on s'en sort ?

Comme pour le point 2), on utilise directement l'hôte ESX qui héberge cette VM pour avoir une console sur la VM, indépendamment du réseau.

6 Validation des différents points de très haute disponibilité

1) Coupure d'un des chemins fibre pour un socle ESX vers la baie SAN.

   On coupe successivement chacun des quatre liens redondants du socle vers la baie SAN (on en laisse toujours 3 actifs).
   Le socle reconnait bien chaque lien coupé, et bascule au besoin sur un autre chemin pour accèder à la baie SAN. Il y a une légère dégradation dans la vitesse d'écriture sur disque pendant la coupure, de quelques secondes. Pas de perte de données.

2) Bascule des liens réseau ethernet physiques pour chaque socle ESX.

   Chaque socle à 4 interfaces réseau physiques. Les interfaces 2 et 3 sont utilisées en interne entre les deux socles pour HA, FT et la migration d'une VM d'un socle à l'autre. Les interfaces 1 et 4 sont utilisées pour faire communiquer l'extérieur avec les socles, mais aussi avec les VMs. Il est donc important d'avoir une redondance réseau pour l'accès aux VMs et aux socles. Par défaut l'interface 1 est en mode actif, et la 4 en mode veille ("standby"). Si le lien avec l'interface 1 tombe, c'est l'interface 4 qui prend la relève et devient active. Il y a 2 à 3s de latence réseau, mais les connexions TCP actives ne sont pas perdues.
   Au retour de l'interface 1, on peut choisir de garder le lien via l'interface 4 ou bien de rebasculer via l'interface 1, le choix actuel étant : rester sur l'interface de basculement.

3) Migration manuelle d'une VM depuis un socle vers l'autre (VM en mode HA).

   OK. Durée < 15s, sans perte de données ni aucune coupure de service.

4) Crash complet d'un socle ESX hébergeant des VM en mode HA+FT
   (coupure alimentation directe).

   OK. Comme indiqué ci-dessus pour le mode 2, pas de coupure de service, un ping seulement est perdu, les connexions TCP sont conservées.
   Lorsque le socle est remis sous tension, il est ré-intégré au cluster automatiquement. De plus le statut des VM initialement sur ce socle repasse automatiquement en mode protégé.

5) Crash complet du VCenter : extinction pure et simple de la machine hébergeant le VCenter
   (machine physique ou virtuelle).

   Le cluster et les VMs continuent de fonctionner normalement, même en mode HA+FT.

6) Ajout de RAM virtuelle à chaud.

   La VM étant en cours de fonctionnement, on augmente sa capacité en RAM via l'interface VCenter (dans les limites respectant HA ou HA+FT disponibles).
   OK, l'OS Windows 2008 sait reconnaître à chaud l'augmentation de RAM, et utilise la nouvelle taille.

7) Ajout de RAM physique à chaud.

   On désactive temporairement le mode protégé (les VMs restant actives), et on migre toutes les VMs sur un socle. On fait l'upgrade RAM physique sur l'autre socle que l'on a préalablement arrêté. Une fois fait on le reboote, ce socle est ré-intégré au cluster, avec sa RAM upgradée. On peut alors migrer toutes les VMs dessus, et faire la même opération d'upgrade de RAM physique sur le premier socle. Une fois fait, on peut remettre le mode protégé pour les VMs. Il n'y a eu aucune coupure de service.

8) Manipulations des disques (virtuels et physiques) à chaud.

A. Mise à jour de LUNs.

On peut détruire/reconstruire des LUNs à chaud (avec les socles ESX actifs), sans avoir à rebooter ces socles : ils reconnaissent à chaud les nouveaux LUNs.
Ceci est notamment très utile pour ajouter à une VM (Windows 2008) un nouveau disque "D :", il suffit de lui ajouter virtuellement via VCenter un "disque dur" supplémentaire, que l'on aura au préalable affecté au nouveau LUN créé. Ceci se fait avec la VM active, sans avoir à la rebooter.
Idem pour la suppression d'un tel "D :".

B. Diminution taille de disque.

On ne peut pas diminuer la taille d'un LUN, ni diminuer la taille d'une partition VMFS sur laquelle repose les fichiers d'une VM. Donc pour diminuer à chaud la taille d'un disque (virtuel) d'une VM, il faut procéder comme suit :

- via l'OS Windows 2008 de la VM, on diminue sa taille du volume, (il fixe le maximum possible à réduire lui-même),
- on migre la VM sur un autre LUN nouvellement créé de taille plus petite, (sans interruption de service),
- on supprime le LUN sur lequel se trouvait initialement la VM.

C. Augmentation taille de disque.

Méthode 1. (VM active, sans interruption de service, durée : environ 1 minute).

- on augmente à chaud la taille d'un LUN attribué à une VM,
- on réanalyse les LUNs via le socle ESX, qui reconnaît l'augmentation de taille du LUN,
- via le Windows 2008 de la VM, on fait réanalyser le volume, il reconnaît l'augmentation, et on peut étendre le volume NTFS à la nouvelle taille attribuée.

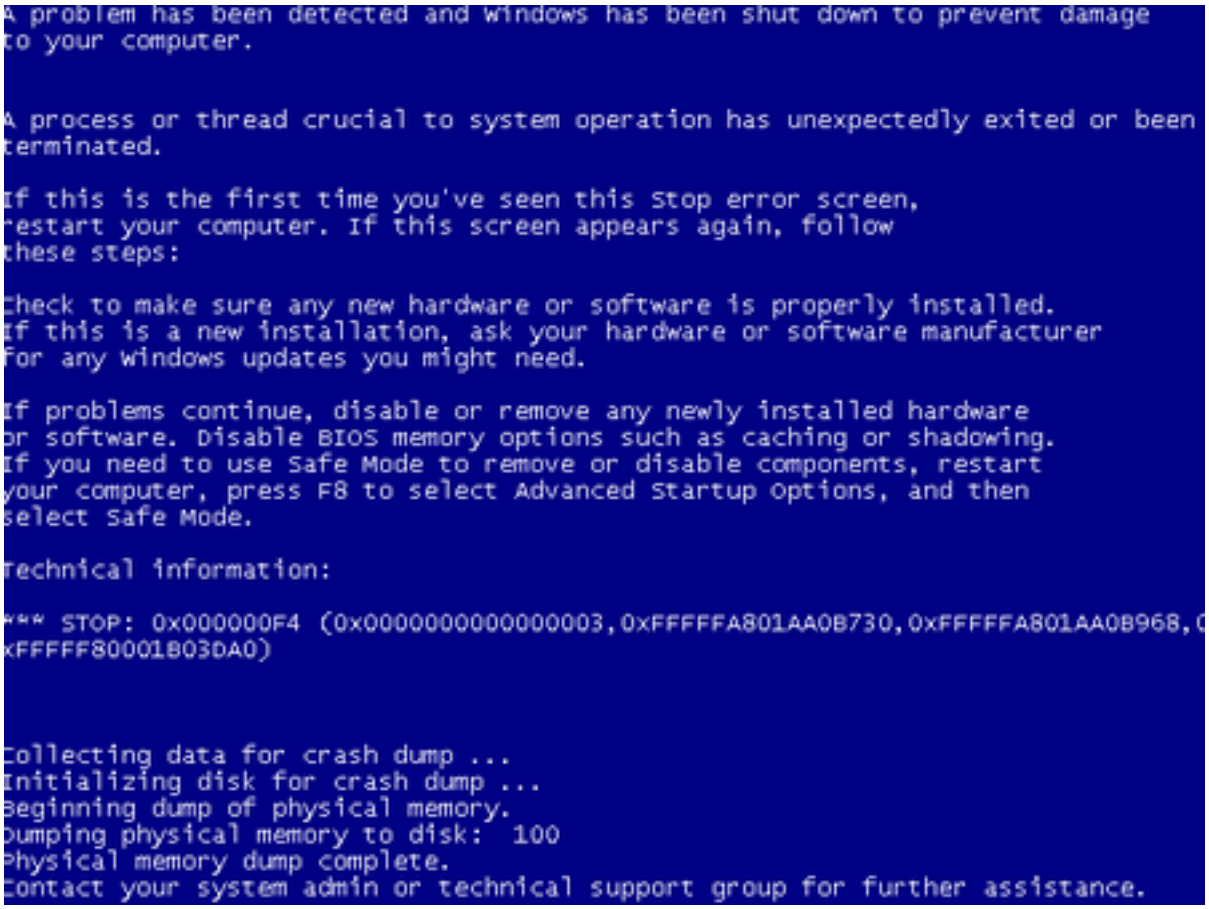
Florian ne recommande pas cette méthode, car l'augmentation à chaud de la taille d'un LUN nécessite sa fragmentation. Il vaut mieux soit avoir fait un SAN design propre dès le début, soit procéder comme suit :

Méthode 2. (VM active, sans interruption de service, durée : environ 5 minutes).

- on crée un nouveau LUN de taille plus grande,
- on migre "à chaud" la VM sur ce nouveau LUN (environ 5 minutes),
- on supprime le précédent LUN devenu inutile.

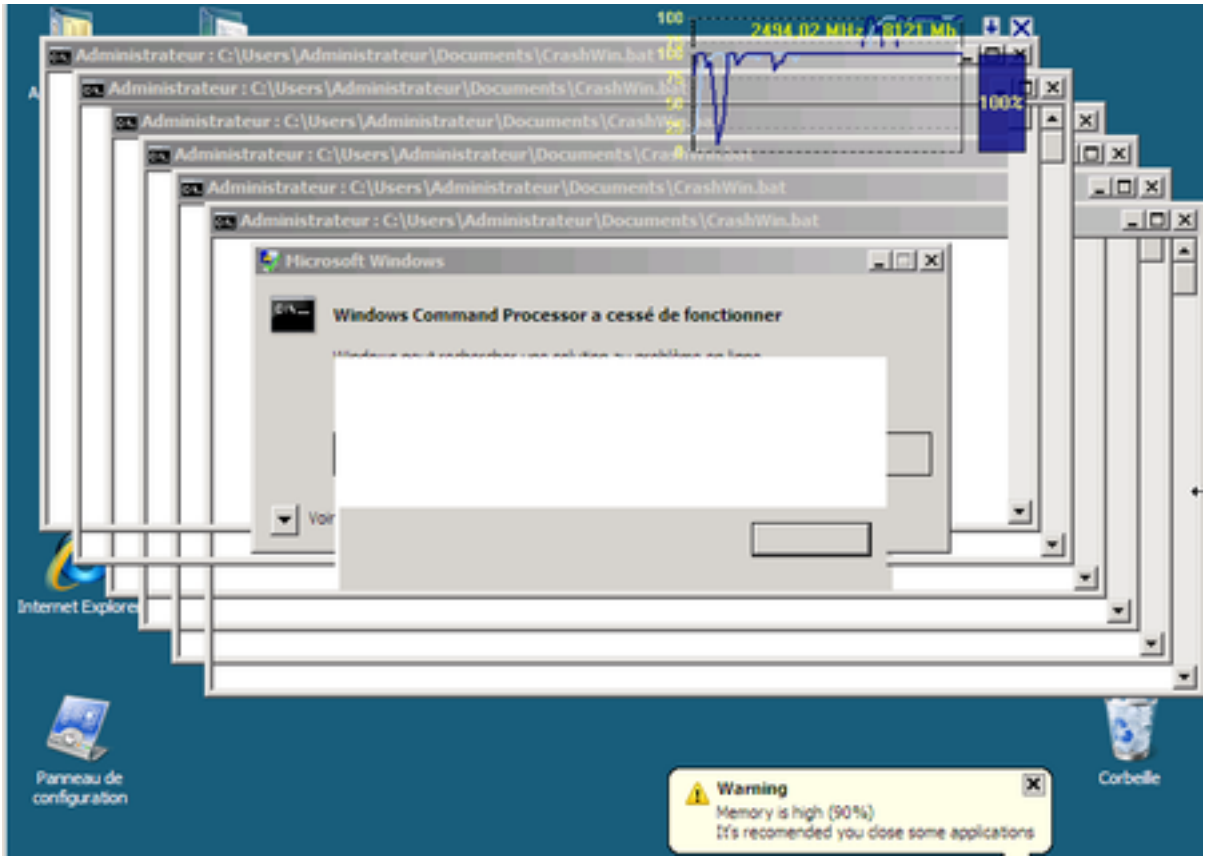
10) Crash tests de l'OS Windows 2008 Datacenter (avec une VM de test).

a) Obtention d'un écran bleu (en local ou bien à distance).



La surveillance de VM via le cluster fonctionne : reboot automatique dans les 15s (durée configurable).

b) Saturation RAM, CPU : Windows rendu inutilisable.



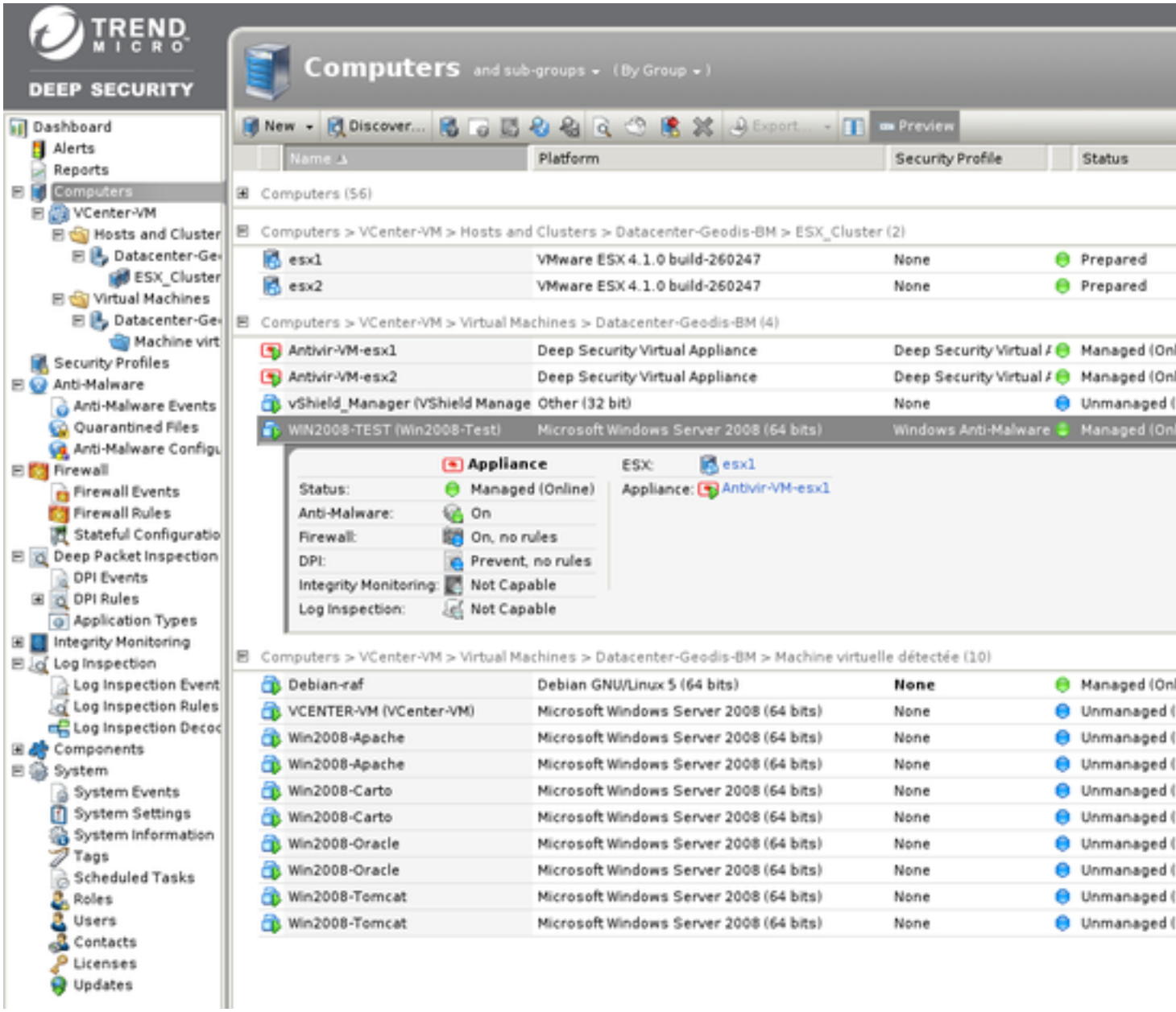
Ici plus d'accès TSE, ni en console. CPU : 100%, RAM : 100%.
Des alertes sont remontées au niveau du VCenter.
Ici la surveillance de VM via le cluster ne fait rien. Il faut probablement se tourner vers d'autres solutions pour avoir une surveillance applicative (Microsoft MSCS, Veritas cluster, Neverfail, Marathon everRun, etc).
Choix actuel : surveillance pour les 5 Windows : Oracle, Tomcat, Apache, Carto et VCenter.

7 Solution Antivirus VMware - TrendMicro

Objet : réduire la charge des OS Windows (processeurs, RAM) dûe à l'utilisation d'antivirus sur chaque VM.

Concept : dédier une seule VM à l'analyse malware (la SVM : Secure Virtual Machine).

Fonctionnement (résumé) : sur une VM protégée, tous les appels en lecture à un fichier (local ou réseau), passent à travers un driver Windows bas-niveau qui envoie les infos du fichier (tout le fichier ?) via le socle ESX à la SVM de ce socle. Selon ce que la SVM décide, le fichier est accepté ou refusé, et le cas échéant supprimé.
Une fois tout installé et bien configuré, on a un accès d'admin à la solution ("Deep Security Manager") qui est se présente comme suit :



Restrictions préalables :

- pour l'instant il n'y a que Trend Micro qui a finalisé cette solution en partenariat avac VMware. La partie TrendMicro est nommée "Deep Security", et celle de VMware nommée "VMware VShield". Parmi les éditeurs d'antivirus, TrendMicro est plutôt bien classé, mais pas autant que les ténors (F-Secure, Kaspersky, Symantec,...).
- cette solution est assez lourde (et délicate) à mettre en place : il faut une VM pour le VShield Manager de VMware, et une SVM TrendMicro dédiée par socle ESX. Il faut en outre installer un driver Windows par VM que l'on veut protéger. Enfin il y a un certain nombre de configurations DNS (ici en local via fichier lmhosts sous Windows et /etc/hosts sous Linux), ainsi que des confs réseau et FW (permettres certains ports et flux entre les VMs protégées, les SVMs et la VShield Manager).

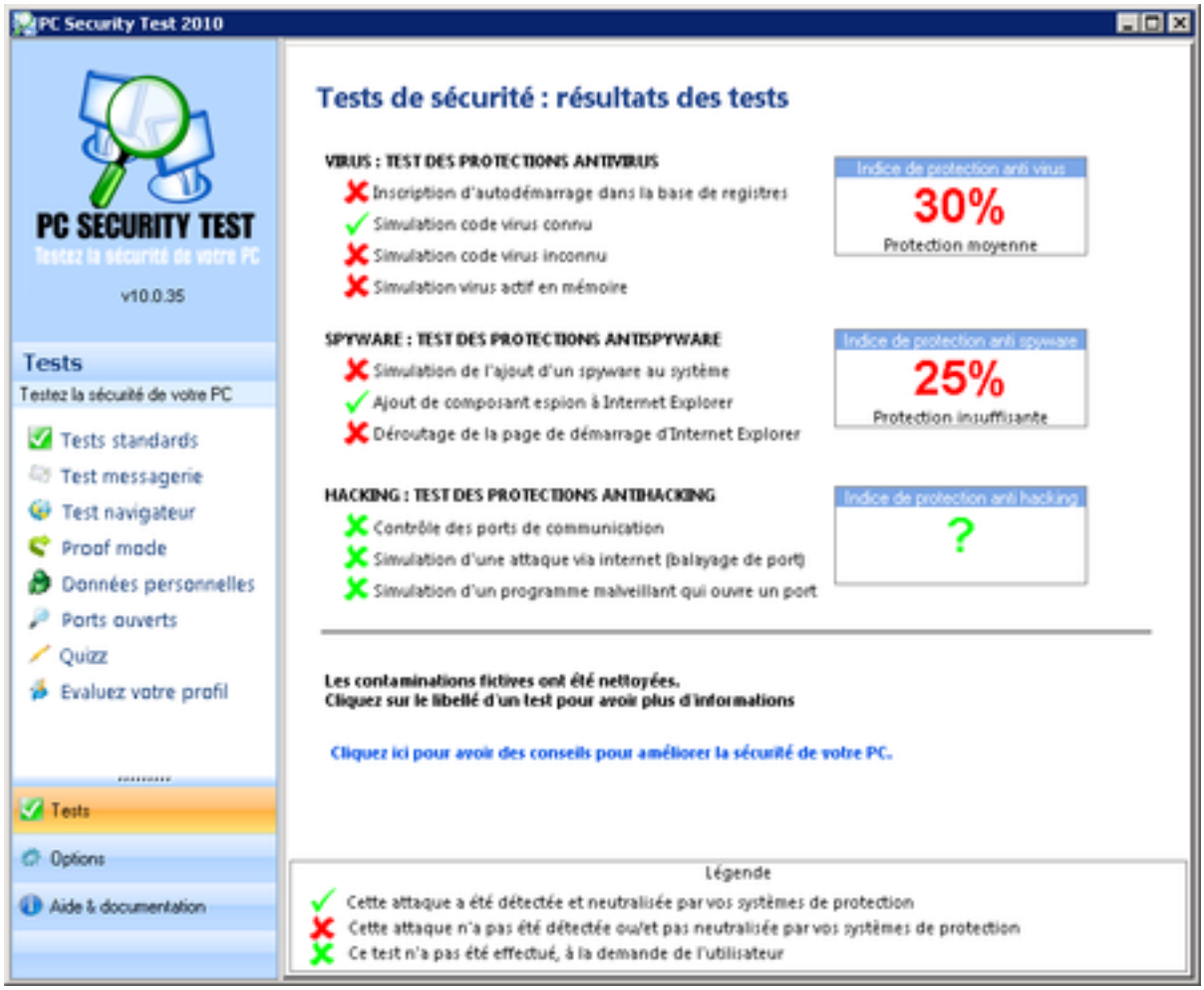
Tests effectués :

1) Installation de virus sur une VM Win2008-Test protégée :

Méthode : je copie directement par transfert de fichiers un virus (virus relativement ancien, pris aléatoirement dans mon stock de virus/malwares).
Résultat : dès l'accès au répertoire contenant ce fichier, le fichier de virus est automatiquement supprimé du disque de la VM, et il est "mis en quarantaine" au niveau de la SVM (on peut le récupérer sur son PC dans un format chiffré et le déchiffrer via un outil spécifique). On voit dans la partie Anti-malware Events du Manager le type de virus incriminé, et le lien vers la description du virus, les éventuelles actions à mener (lien directement sur le site web de TrendMicro).

2) PC Security Test 2010.

Ce programme simule des attaques virales comme indiqué ci-dessous :



Ce test est un peu biaisé, car ceux qui le proposent ont une solution anti-virale à vendre.

Ce qui est clair, c'est que la solution VMware-Trend ne fonctionne que sur les fichiers, et pas sur des virus activés en mémoire directement (par exemple via un contrôle ActiveX...).
De plus cette solution ne protège pas contre les écritures malicieuses dans la base de registre, comme par exemple via un simple script dos.bat (testé).

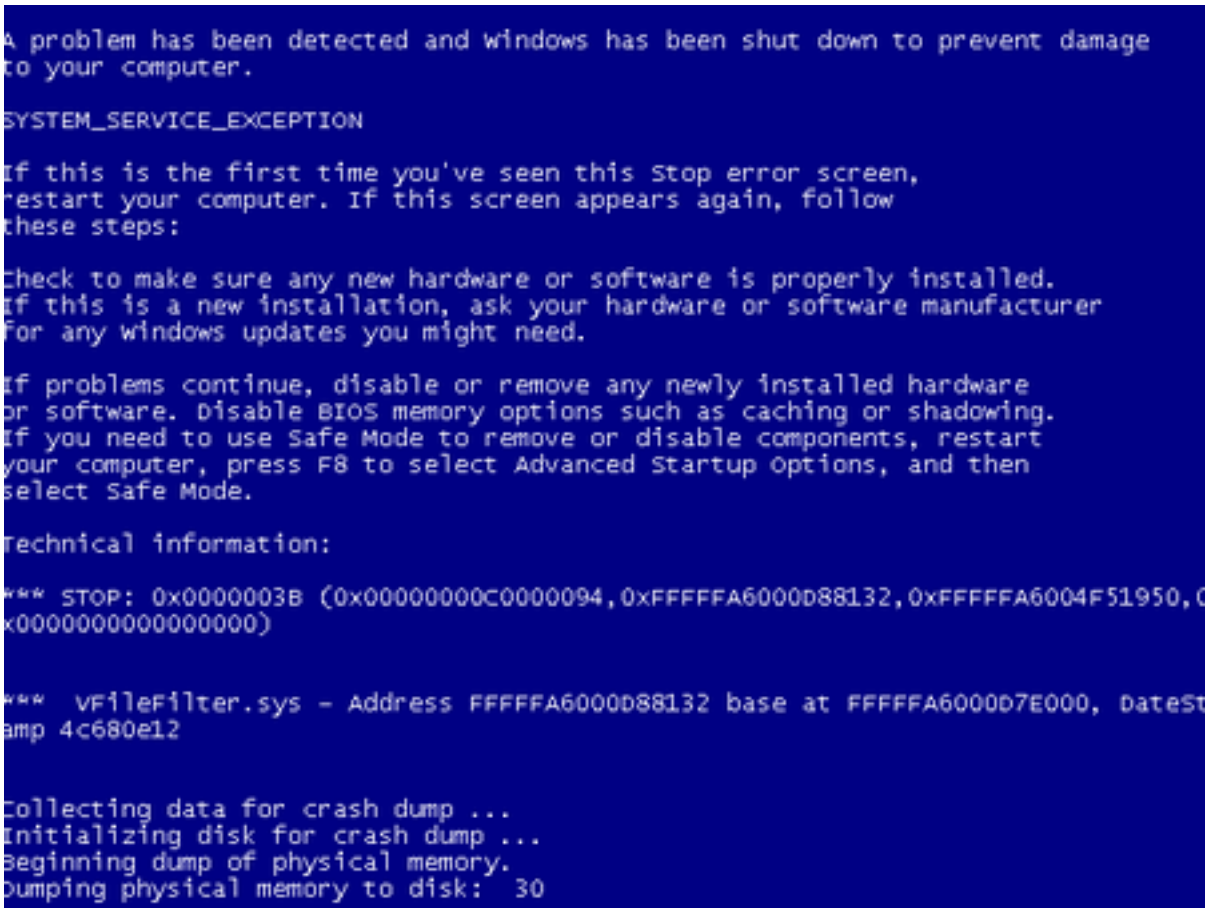
3) Tests Kaspersky Lab (http ://support.kaspersky.com/fr/viruses/avtest)

`suspicious.exe` : test fichier suspect (code virus "inconnu").
Résultat : il a été supprimé, reconnu en "TROJ_LAMEWAR.VTG".

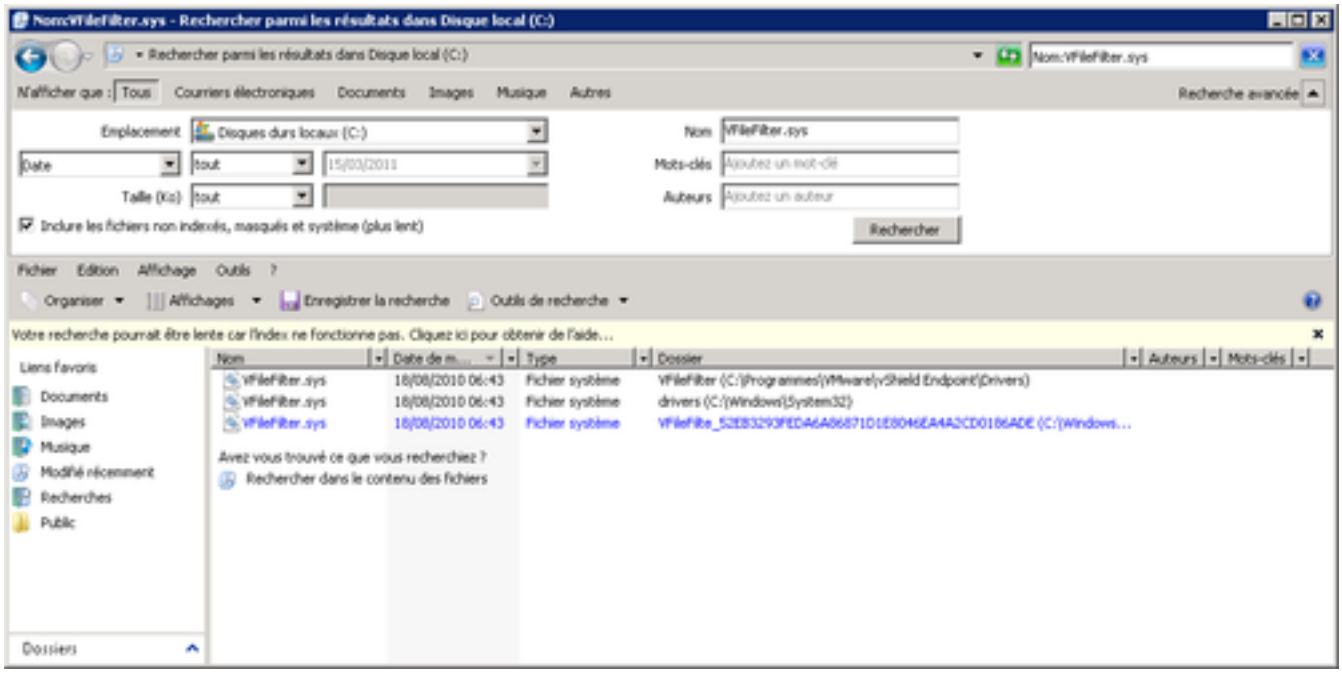
`warning.exe` : code modifié d'un virus connu,
`cured.exe` : corps texte du virus est remplacé par le mot "DISINFECTED",
`deleted.exe` : ???
Résultat : pour ces 3 derniers, comme ce sont des exécutables 32 bits, (et que l'OS est Windows 2008 en 64 bits), ils ont été laissés en place sans modifications.

4) La découverte du gros BUG.

Il faut dire que cette solution est conceptuellement très intéressante et très avantageuse. Mais son défaut : elle vient de sortir, et du coup est proche de la version beta. En outre : le driver bas-niveau Windows est fait ... pour Windows, ce qui pose tous les problèmes de ... Windows !
Le bug a été obtenu en faisant un test d'injection de virus par un partage Windows (comme un partage Samba).
En manipulant de façon un peu tordue les autorisations des utilisateurs d'un partage sur une machine Windows, tout autre VM Windows protégée par VShield - Deep Security, qui tenterait d'exécuter un binaire se trouvant sur ce partage, se retrouve dans l'état suivant :



Ce qui est intéressant dans cet écran bleu, c'est la ligne concernant `VFileFilter.sys`.
C'est le driver bas-niveau de VMware VShield Endpoint :



On le retrouve dans la base de registre au niveau des clés :
`HKEY_LOCAL_MACHINE\SYSTEM\ControlSet001\Services\VFileFilter`

Je pense qu'un programme malveillant (je dois pouvoir le faire), peut très bien effectuer l'une ou l'autre des tâches suivantes :

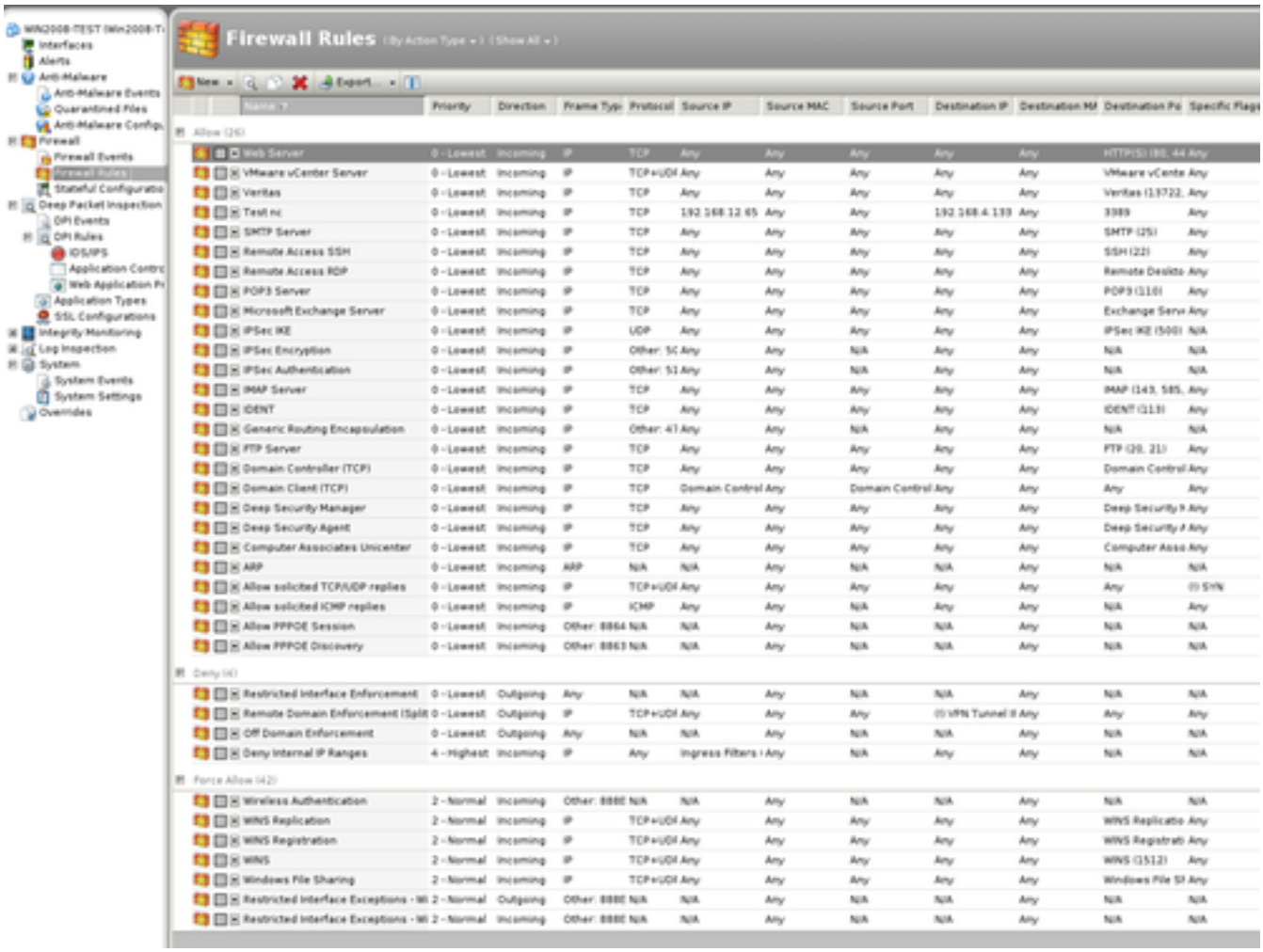
- désinstaller le driver VMware VShield Endpoint (puisqu'on peut le faire manuellement),
- supprimer les fichiers type `VFileFilter.sys` dans les répertoires `C:\Windows\System32` et `C:\Programmes\VMware\VShield Endpoint\Drivers\`,
- supprimer les clés de registre en rapport avec `VFileFilter.sys`, notamment son démarrage automatique,

Je n'ai pas trouvé sur le site de TrendMicro de sujets en rapport à ce bug. Peut-être faut-il prendre le temps de leur signaler ?

5) Sécurité et performances

- Si la sécurité anti-virale est très importante, outre le BUG ci-dessus (qui sera a priori réparé), il y a 2 points critiques : la VM VShield Manager, et la SVM d'un socle. Si l'une de ces deux machines a un problème qui l'empêche de communiquer avec les autres, il n'y a plus de sécurité pour les VMs du socle. A noter quand même que ces machines ont pour OS Linux, ce qui inspire un peu plus confiance (en tout cas à moi).
- Il manque les tests de montée en charge au niveau IOs d'écriture sur disque avec (et sans, pour comparer) cette protection antivirus d'activée.
  De bons benchmarks ont été faits ici :
  http ://geeksilver.wordpress.com/2010/12/17/vmware-vshield-endpoint-and-trend-micro-deep-security-7-5-understanding-part-3/

- A noter qu'il y a une bonne protection FW (au niveau des socles ESX) pour chacune des VMs, qui peut être faite via l'interface d'admin Deep Security. Il y a des templates de règles de FW que l'on peut appliquer aux VMs selon leur profil, et l'on peut ajouter/supprimer des règles ponctuellement en fonction des besoins. Voici un aperçu :



- A noter aussi que l'installation du driver VShield Endpoint sur les VMs que l'on souhaite protéger entraîne forcément le reboot de celles-ci.

8 Divers (sur le cluster VMware ESX)

1) Outre un suivi live des performances CPU/RAM/disques des VMs du cluster ainsi que des socles le constituant, on peut exporter un historique de ces performances au format excel.
2) Testé : l'importation (copie/conversion) d'une machine physique en une VMs. En moins de 15 minutes, j'obtiens une copie de la machine physique sous Windows 2008 Datacenter qui hébergait initialement VCenter en une VM fonctionnelle.
   Note : la copie de Windows apparaît alors comme "copie non authentique" et doit être activée dans les 30 jours, d'où le choix de refaire une VM Windows 2008 avec VCenter mais en version d'évaluation 240 jours.

3) Testé : le clone d'une VM.
Idem, obtenu en moins de 15 minutes, pour un disque de 50 GO.

4) La solution VCenter sous Linux n'est pour l'instant pas envisageable : il n'existe pour l'instant que la version 2.5 (d'où des problèmes de compatibilité avec les socles ESX 4.1), et elle n'a pas toutes les fonctionnalités de configuration des VMs, de plus est très délicate à mettre en oeuvre : http ://communities.vmware.com/docs/DOC-9580
Une version 4.x pour Linux est en cours de développement, les accords commerciaux VMware-Microsoft freinent le processus (http ://communities.vmware.com/thread/219255 ?tstart=0).

9 Solutions concurrentes

1) KVM/Qemu

Gratuit (GPL) mais supporte beaucoup moins de fonctionnalités, et pas du tout le mode HA, encore moins HA + FT.
http ://www.linux-kvm.org/page/Main_Page

2) RedHat Entreprise Virtualization (optimisation payante de KVM/Qemu).

Depuis peu sur le marché, moins cher que VMware, supporte le mode HA, mais pas encore le mode HA + FT.
Globalement nettement moins "mûr" que VMware vSphere au niveau des fonctionnalités.
http ://communities.vmware.com/message/1656474#1656474
http ://myvirtualcloud.net/p=526&utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A

3) Xen (Open Source, GPL)

Produit qui ne supporte pas bien Windows, au mieux Windows 2003. En comparaison avec VMware vSphere, c'est un logiciel en version infra-beta...
http ://fr.wikipedia.org/wiki/Xen#cite_note-1
http ://wiki.xensource.com/xenwiki/XenFaq

4) XenServer (Citrix), Hyper-V (Microsoft), etc

Solutions commerciales que je ne connais pas.