Retour d’expérience suite à une défaillance du stockage de l’infra de virtualisation

Frédéric Pauget
Télécom Paris
19, place Marguerite Perey
91 123 Palaiseau

Résumé

Début 2021, en pleine nuit du vendredi au samedi, la baie de stockage hébergeant les disques de la majorité des machines virtuelles de Télécom Paris plante. En un instant, on bascule de « tous les services fonctionnent » à « rien ne marche ». Une situation redoutée dans n’importe quelle DSI. Une baie de disque en panne, c’est un peu une grosse boite noire autour de laquelle on peut prier pour que le support du constructeur la remette en service rapidement. C’est aussi un moment de doute : « est-ce que la mise en œuvre de mon PRA va se dérouler sans accroc ? »

Cet article décriera en première partie le contexte et les infrastructures, puis, dans la seconde partie, le déroulement de l’incident jusqu’à sa résolution, et enfin la dernière partie donnera les enseignements que nous avons pu tirer de cet évènement.

Mots-clefs

PRA

1 Le contexte

Télécom Paris est une école d’ingénieurs comprenant 1600 étudiants, 220 doctorants et 161 enseignants chercheurs. Elle fait partie de l’IMT (Institut Mines Télécom) et de l’Institut Polytechnique de Paris.

1.1 Les services

Les principaux services fournis par la DSI sont :

- le réseau sur site et les services de base associés (DNS, DHCP, VPN),
- le stockage de fichiers pour les personnels de l’école et de la Direction Générale de l’IMT,
- la messagerie de 4 écoles et de la DG IMT et IP Paris (14 000 boites),
- le SI scolarité pour 7 institutions,
- l’hébergement des SI finance et RH de l’IMT,
- des hébergements Web et de diverses applications,
- des services d’authentification pour les usagers gérés : LDAP, Kerberos, Active Directory, fournisseur d’identité SAML.

1.2 L'infrastructure

1.2.1 L'architecture de production

L'architecture de production est intégralement déployée sur le site de Palaiseau dans une unique salle serveur.

La plupart des services sont hébergés par un cluster de virtualisation Promox, constitué de 15 nœuds permettant l'exécution de 220 machines virtuelles. La très grande majorité de ces VM utilise un stockage NFS, sur une baie DELL EMC Unity 550F, pour un volume d'environ 80To. C'est une baie composée de 25 disques flash en RAID, deux contrôleurs redondants et deux alimentations électriques également redondantes.

L'interconnexion entre la baie de stockage et les nœuds de virtualisation est assurée par un réseau redondant à base de connexions 10Gb.

Cette architecture à base de stockage partagé comporte plusieurs avantages :

- ne pas lier la capacité de stockage aux capacités CPU/RAM et donc optimiser l'allocation des ressources au mieux,
- une migration des VM à chaud entre les nœuds et donc ainsi assurer une meilleure continuité de services lors de besoins de maintenances matérielles ou logicielles sur les serveurs physiques,
- l'utilisation d'instantanés sur l'intégralité des machines virtuelles,
- des capacités de compression et déduplication efficaces (43 %).

Elle a aussi un inconvénient : la baie constitue un point de défaillance unique (SPOF). Toutefois, le haut niveau de redondance de celle-ci doit permettre d'assurer une disponibilité optimale.

Les quelques VM qui utilisent les disques locaux des serveurs sont celles hébergeant des services redondants, l'arrêt du nœud n'impactant alors pas la production. Les services concernés sont le DNS, le DHCP, le LDAP, le Kerberos, l'Active Directory et le fournisseur d'identité SAML.

Enfin quelques services tels que les répartiteurs de charge et le VPN ne sont pas virtualisés.

1.2.2 L'architecture de sauvegarde – reprise d'activité

La baie principale Unity 550F du site de Palaiseau est répliquée de manière asynchrone sur une baie Unity 330 située sur le site de Télécom Sud-Paris à Evry. L'intervalle de réplication est de une heure.

Contrairement à la baie de production possédant uniquement des disques flash, cette baie de réplication est composée de disques SATA 7200rpm ainsi que de quelques disques flash de cache.

Cette baie sert aussi de sauvegarde pour les machines virtuelles hébergées localement sur les nœuds de virtualisation.

Point de vue réseau, cette baie est reliée par deux liens gigabit. L'interconnexion entre les sites de Palaiseau et d'Evry est réalisée via un site parisien, avec une liaison optique à 10Gb. Bien que permettant très facilement l'acheminement de liaisons L2 via l'utilisation de VLAN, nous avons préféré garder des adressages IP disjoints pour les deux baies et donc passer par les infrastructures de routage de nos deux écoles. Ce choix a été fait afin d'assurer la redondance : en cas de défaillance de la liaison dédiée, le trafic passe automatiquement par le L3VPN sur Renater mis en place pour l'IMT.

2 L'incident

2.1 Signes précurseurs

Courant Juin 2020, un disque de la baie de production indique une fin de vie proche tout en continuant à fonctionner normalement. Cela est étrange, la baie datant seulement de début 2019. Le support Dell est immédiatement informé. Après investigations, le support nous informe que l'information remontée est fausse, que le fonctionnement de notre baie est normal et supprime l'alarme. Cela se reproduit à plusieurs reprises sur d'autres disques avec le même résultat.

Fin 2020, plusieurs disques sont dans cet état et le support nous informe que le problème est réglé dans la nouvelle version du micrologiciel et qu'une mise à jour de celui-ci réglera le problème.

2.2 Samedi matin

Début 2021, dans la nuit du vendredi au samedi, la plupart des services plantent. Malgré l'absence d'astreinte à la DSI, l'information « il y a plus rien qui marche » est remontée rapidement jusqu'à une personne gérant l'infrastructure. En connexion distante, le diagnostic est rapide : le NFS sur lequel repose la virtualisation est hors service. L'investigation continue sur l'interface d'administration de la baie : le pool de disque est hors ligne.

Le support du constructeur est alors alerté. Grâce à une prise en main à distance via une session Zoom, il peut travailler sur la baie. Les premiers éléments n'indiquent aucune défaillance physique. A priori, il est donc possible de la remettre en service. Les données ne sont pas perdues mais doivent être reconstruites. Par contre, le temps de reconstruction des données ne peut être évalué.

Afin de nous rassurer sur nos données, nous vérifions l'état de la baie de reprise d'activité : fonctionnement indiqué optimal et données synchronisées jusqu'à l'incident.

2.3 Samedi soir

La reconstruction est toujours en cours, nous nous fixons pour objectif une reprise d'activité au plus tard le lundi matin, avant la reprise d'activité des usagers. Deux scenarii :

- la reconstruction se termine suffisamment tôt et on reprend le service sur l'infrastructure normale,
- elle prend trop de temps et on utilise la baie de reprise d'activité située sur le site d'Evry.

Le premier scénario est privilégié. Il se pose toutefois la question de savoir pourquoi ce plantage est survenu. Le second entraîne beaucoup plus d'interrogations : est-ce que la baie et le réseau vont tenir la charge ? Pour le réseau, les statiques de débit indiquent que ce ne sera pas un problème mais quid de la latence : de l'ordre de 0,1 ms on passera à 0,5 ms. L'inquiétude principale reste la baie en elle-même. En théorie, la baie d'Evry possède des performances relativement équivalentes à l'ancienne baie de production, remplacée fin 2019 par la baie Unity 550F. Cela devrait donc fonctionner avec des performances dégradées, identiques à celles de deux ans en arrière. Enfin dans l'hypothèse d'utilisation de la baie d'Evry, une opération de bascule arrière sera nécessaire, une fois la baie principale remise en service.

Le temps de remise en service sur la baie d'Evry est estimé à deux heures. Nous décidons donc d'attendre le dimanche soir pour trancher.

2.4 Dimanche fin d'après-midi

La reconstruction s'avère être une succession de plusieurs étapes, ayant chacune un pourcentage de progression, des vitesses variables. Il est donc vraiment impossible d'estimer le temps de reconstruction des données. Les techniciens du support nous indiquent toutefois qu'il est peu probable que cela soit fini avant lundi.

Nous prenons la décision de rétablir les services en activant notre PRA, sur la baie d'Evry, plan de reprise d'activité élaboré de longue date, pour palier ce type de problème.

« bah si il y a un problème on utilise l'autre baie »

Figure 1: Plan de reprise d'activité (version intégrale)

Avec l'aide du support, les données répliquées sont promues source et le service NFS est activé. Les réseaux étant différents, une petite manipulation est nécessaire pour spécifier la nouvelle IP de service, sinon la baie aurait automatiquement repris l'IP de la source. Le cluster de virtualisation est également reconfiguré pour pointer sur la nouvelle IP. Cette première phase prend moins d'une demi-heure, utilisée surtout pour se familiariser avec les subtilités de l'interface de la baie pour ce type de manipulation. Enfin, nous relançons les machines virtuelles en privilégiant celles jugées indispensables. Ce choix de ne pas relancer toutes les machines virtuelles est dicté par les incertitudes sur la tenue de charge. Les machines « éliminées » sont principalement celles des infrastructures de développement et pré-production.

Moins de trois heures après la prise de décision de basculer sur le PRA, les services sélectionnés sont complètement opérationnels. Il apparaît bien compliqué de démarrer les machines virtuelles unitairement, la fonction de démarrage en masse de notre hyperviseur ne pouvant être utilisée dans ce processus de sélection.

2.5 Les semaines suivantes

Dès le lundi matin, les usagers ont accès à la plupart de leurs outils informatiques habituels. Bien que dégradées, les performances restent acceptables. Au cours de la matinée, nous redémarrons de plus en plus de machines virtuelles jusqu'à un fonctionnement complet de l'ensemble des services.

Pendant ce temps, la reconstruction de la baie principale, sur la site de Palaiseau, se poursuit. Elle se terminera le mardi, en fin d'après-midi. Une fois terminée cette étape, la réplication est immédiatement remise en service, dans le sens Evry vers Palaiseau. Celle-ci nous permettra la bascule inverse. Le jeudi, les deux baies sont totalement synchronisées.

Il faut maintenant penser à revenir en situation nominale. Cependant, avant de remettre la baie en production, nous souhaitons connaître la cause de la défaillance. Malgré nos demandes insistantes, la seule explication qui nous sera fournie est que le micrologiciel était trop ancien, que le problème provenait d'un bug corrigé dans la dernière version.

Les diverses mises à jour logicielles sont appliquées avant la reprise de la production sur la baie principale. Dans l'objectif de limiter les impacts sur le service, nous souhaitons mettre à jour la baie SSD qui est à ce moment hors production, vérifier que la réplication se déroule correctement puis basculer le service sur celle-ci (coupure). Le problème est que le support n'a pas d'information fiable sur le fonctionnement possible de la réplication entre des baies de version logicielle différentes !

N'ayant pas trop de choix, nous décidons de mettre en œuvre ce scénario ; si la réplication échoue, une nouvelle coupure de service sera nécessaire pour l'upgrade de la baie secondaire. En définitive,

la réplication fonctionne bien après la mise à jour et donc nous pouvons maintenir le plan prévu initialement. Celui-ci se déroule sans accroc avec moins d'une heure de coupure de service. Deux semaines après l'incident, nous revenons en situation nominale.

3 Les enseignements

Dans chaque chose que nous faisons, il y a un enseignement à tirer.

3.1 Le Plan de Reprise d'Activité

Cet incident nous a permis de valider notre PRA bien qu'il tienne sur un ticket de métro : une bonne maîtrise de l'infrastructure permet de mettre en place les bonnes opérations pour relancer les services.

Néanmoins ne l'ayant pas testé grandeur nature, nous avons longuement hésité à activer le PRA, en raison des incertitudes sur la charge. A posteriori, il apparaît que cela aurait été bénéfique de l'activer dès le samedi. Si le problème s'était déclenché en semaine, il aurait eu plus d'impact sur les usagers, mais nous n'aurions pas attendu aussi longtemps.

Pour le futur, cet évènement passé nous donne des éléments mais pas la réponse à la question « espérer une réparation rapide ou appliquer le PRA », cela dépend de l'incident.

3.2 Les infrastructures

Nous avons été confortés dans notre choix d'architecture : des services d'infrastructure de base (notamment DNS, LDAP et VPN) hautement disponibles, non soumis à un SPOF. C'est grâce à la disponibilité de ceux-ci qu'il a été possible d'intervenir rapidement à distance. Il aurait été beaucoup plus laborieux de devoir se déplacer sur site et se retrouver avec un réseau dysfonctionnel.

La réplication totale du stockage est aussi un point important qui nous a permis d'éviter d'attendre le mardi pour pouvoir relancer les services, et surtout nous a rassuré dès la survenue de la panne : savoir que, en cas d'échec de la réparation de la baie, les moyens de remettre l'infrastructure en place sans avoir perdu de données existent est sécurisant.

Avoir la baie secondaire sur un autre réseau complexifie légèrement les opérations de reprise avec le besoin de reconfigurer les IP sur la virtualisation et peut entraîner une légère dégradation des performances, mais cela n'a pas été problématique. Mais là encore il y a un côté rassurant à avoir plusieurs chemins réseau quand la plupart des services sont dépendants d'une connexion réseau.

Suite à cet incident, nous avons toutefois décidé de faire évoluer l'infrastructure en faisant un pas vers la continuité de service en ajoutant une seconde baie en réplication synchrone avec la baie principale. Elle est hébergée dans la même salle serveur que de la baie principale, est également équipée de disques SSD, mais sa capacité est limitée aux services importants. La réplication synchrone permettra en théorie de passer d'une baie à l'autre sans avoir à redémarrer les machines virtuelles. Une réplication asynchrone est conservée avec le site d'Evry.

Enfin cet incident nous aura alerté sur un point important : la gestion des versions de micrologiciels. Ce n'est pas parce que le système fonctionne et donne satisfaction depuis plusieurs mois qu'il faut négliger les mises à jour. Une mise à jour, c'est le risque de provoquer une interruption de service mais une interruption programmée et maîtrisée.

3.3 La communication

On peut distinguer plusieurs besoins de communiquer

3.3.1 Avec le support technique du constructeur

L'engagement du support se fait assez simplement par téléphone (numéro sur le site public) avec uniquement le numéro de série de la baie. Ces informations sont faciles à trouver, mais cela aurait pu être plus compliqué s'il fallait un numéro de client ou des identifiants à un site spécifique de support. Cela est un point à vérifier dans la conception d'un PRA.

Une fois contactés, les supports des constructeurs ont généralement besoin d'un outil de prise en main à distance, et si on ne dispose pas ce celui-ci il faut l'installer. Dans le cas de Dell, Zoom a été utilisé, outil installé et bien connu au sein de notre école. Là encore dans la conception d'un PRA, il est mieux de ne pas avoir à faire des paramétrages de dernière minute sur un poste de travail.

Les supports constructeurs travaillent 24h/24 ; quand un technicien a fini sa journée de travail, un autre prend le relais (souvent dans un autre fuseau horaire). Notre DSI n'est cependant pas organisée comme cela (comme la plupart des DSI de nos institutions académiques). Dans notre cas, la session de support a été lancée à distance via le VPN sur un poste de travail à domicile sans avoir anticipé que l'incident allait durer aussi longtemps. Cela a monopolisé l'ordinateur et a empêché la prise de relais par un autre collègue. Il aurait peut-être été possible de couper la session pour la reprendre, mais nous avons préféré ne pas prendre ce risque, la connexion a donc été maintenue les quatre jours. Cela étant assez peu pratique, pour le futur nous avons mis en place une machine dédiée sur site sur laquelle il est possible de prendre la main à distance en se partageant une session.

3.3.2 Au sein de la DSI

Le traitement de l'incident ne demandant pas beaucoup de main d'œuvre de la DSI, le besoin de communication interne entre les différentes personnes pouvant intervenir a été assez faible et on a pu se débrouiller par téléphone et SMS. On peut distinguer deux canaux : un en direction de l'équipe de l'infrastructure : « c'est en panne, c'est géré, si besoin je te rappelle » et un autre vers le DSI pour des points de situation réguliers en fonction de la reconstruction et de prendre la décision de basculer la production.

3.3.3 En direction des usagers

La plupart des systèmes étant hors service, il était impossible d'utiliser les canaux usuels de communication (courriel, sites web de l'école) pour indiquer l'incident aux usagers. Pour ce point nous n'avons pas identifié de moyen idéal. Le problème s'étant produit un week-end et les services étant opérationnels sur notre PRA le lundi matin, cela a réduit le nombre de personnes impactées. Notre DSI a toutefois communiqué par SMS très régulièrement avec la direction de l'école afin de la tenir informée des évènements. Les DSIs des autres écoles ayant des services hébergés sur notre baie ont été informés également par SMS tout au long de l'incident.

4 Conclusion

Devoir faire face à un incident majeur n'est généralement pas un plaisir, mais finalement on aime quand un plan se déroule sans accroc !