Migration vers une double infrastructure hyperconvergée

Rafael Diaz Maurin

Migration vers une double infrastructure hyperconvergée

Rafael Diaz Maurin
Équipe Systèmes
Pôle Infrastructures
DSI de l'Université de Rennes1

Résumé

La DSI[1] de l'université de Rennes1 opérait depuis 2013 une infrastructure de virtualisation et de stockage architecturée autour d'un SAN[2] classique, connecté à plusieurs hyperviseurs VMware, afin de fournir plus de 500 machines virtuelles. En 2018, notre SAN ne nous convenait plus : performances dégradées, peu de maîtrise par l'équipe, pas de reprise automatique, manque d'élasticité, fin de vie annoncée par l'éditeur.

L'Université de Rennes1 compte plus de 30 000 étudiants et près de 3 500 personnels. La DSI opère plus de 200 services pour les utilisateurs dont l'écrasante majorité tourne sur nos propres infrastructures et de nouveaux projets gourmands en ressources étaient envisagés.

En évaluant les différentes solutions de stockage, nous nous sommes intéressés au stockage distribué et plus particulièrement à Ceph. Nous pensions alors adosser ce stockage aux hyperviseurs vSphere, mais l'intégration de Ceph dans ce type d'architecture s'avère complexe et trop peu performante. Nous avons donc choisi l'hyperviseur opensource KVM[3], qui pouvait tirer parti de Ceph nativement. Conjointement nous avons découvert avec enthousiasme les HCI[4] matérielles, dont les coûts dépassaient toutefois nos budgets. Proxmox Virtual Environment[5] est une HCI logicielle : une solution de virtualisation libre basée sur KVM qui intègre Ceph. En parallèle, nous constatons que le SDS[6] de VMware (vSAN) est devenu plus simple à installer.

C'est pourquoi nous avons fait le choix de déployer nous-mêmes les doubles clusters Proxmox/Ceph et vSphere/vSAN, sur du matériel commun. Ce socle de serveurs identiques nous permet de pouvoir réutiliser les hyperviseurs ou les disques dans l'une ou l'autre des infrastructures. De plus, opérer deux infrastructures nous donne la liberté de choisir la technologie de virtualisation pour les VM[7] hébergées en fonction des contraintes propres à chaque service.

Cet article détaillera les avantages et inconvénients de chaque solution, ainsi que les atouts et les faiblesses du maintien en condition opérationnelle (MCO) de deux technologies concurrentes.

Mots-clefs

HCI, hyper-convergence, virtualisation, stockage, Proxmox, Ceph, KVM, vSphere, vSAN, double infrastructure, migration, renouvellement, investissement, container LXC, NAS

1 Direction du Système d'Information

2 Storage Area Network : réseau de stockage dédié et directement accessible en mode bloc par les serveurs

3 Kernel-based Virtual Machine : hyperviseur libre pour Linux

4 Hyper-Converged Infrastructure : plusieurs serveurs physiques agrègent étroitement des composants de stockage, de réseau et de virtualisation

5 https://www.proxmox.com/en/proxmox-ve

6 Software-Defined Storage : terme marketing pour désigner le stockage de données basé sur des solutions logicielles

7 Machine Virtuelle : logiciel ou système d'exploitation qui affiche le comportement d'un ordinateur physique

1 Introduction

Notre système de stockage principal (baies SAN Compellent SC8000, connectées via des *fabrics Fibre Channel*[8]), utilisé depuis 2013 pour héberger plus de 500 *VM*, arrivait en fin de vie. Nous devions le renouveler. L'exploitation en avait été confiée au constructeur dans le cadre d'un contrat de sous-traitance.

La forte croissance des besoins de volumétrie du stockage et la nécessaire augmentation de la réactivité devenaient difficiles à concilier avec notre modèle. Ces difficultés croissantes nous ont conduits à reconsidérer notre stratégie. L'existence de multiples autres systèmes de stockage annexes nécessitait une rationalisation.

L'équipe souhaitait unifier ces moyens de stockage, en regagner la maîtrise technique et financière, gagner en flexibilité dans la gestion des extensions. Nous avons étudié les *HCI* qui semblaient porter ces promesses.

2 Étude des infrastructures hyper-convergées

Une infrastructure hyper-convergée repose sur une pile logicielle utilisée pour gérer les éléments d'infrastructure (vCPU, vRAM[9]…) dans un groupe de ressources partagées, en intégrant un système de stockage distribué et réparti sur les différents nœuds de virtualisation.

On peut la comparer à une infrastructure de virtualisation intégrant le stockage, devenu hautement disponible pour les *VM*. L'intérêt principal est de permettre la croissance de l'infrastructure de façon linéaire et granulaire sans remettre en cause l'architecture globale.

L'ajout d'un nœud dans le cluster permet d'augmenter les différentes ressources physiques (*scale-out*[10]) : on dispose alors d'une plus grande quantité de vCPU, de vRAM et de stockage disponible. Dans certains cas (Ceph), cela améliore conjointement les performances du stockage en augmentant les I/O, car chaque nœud contrôle physiquement les disques qui y sont attachés.

On peut également augmenter les ressources et les performances en ajoutant des ressources matérielles, de façon homogène, dans chaque nœud existant (*scale-up*[11]).

Les solutions d'hyper-convergence peuvent être de deux types :

– soit des applications matérielles (*appliances*[12]) préconfigurées par les constructeurs,

– soit des solutions purement logicielles dans lesquelles sont associés des logiciels hyperviseurs (vSphere, KVM+ProxmoxVE) à des solutions de stockage à définition logicielle (*Software-Defined-Storage*[13]) comme Ceph et vSAN.

2.1 *HCI logicielles* plutôt qu'*HCI matérielles*

En 2018 quand nous nous sommes intéressés aux solutions *HCI* matérielles, les trois principaux acteurs étaient : Nutanix, HPE Simplivity et Dell/EMC VxRail. Réputés robustes et fiables, ces boîtiers sont séduisants sur le papier, mais leur coût reste très élevé, voire prohibitif.

8 Protocole de stockage par blocs, défini par la norme ANSI X3T11

9 Quote-part d'un processeur ou de la mémoire d'un serveur physique affecté à une machine virtuelle

10 Évolutivité horizontale : augmentation des ressources et des performances par ajout de plus d'équipements (ajout de nœuds)

11 Évolutivité verticale : augmentation des ressources et des performances par l'ajout de matériels (RAM, disques) dans les équipements

12 Équipement informatique associé à un logiciel préinstallé et dédié à une fonctionnalité

13 Le stockage défini par logiciel apporte une abstraction entre le logiciel et le matériel au sein d'une solution de stockage et s'oppose ainsi aux SAN, NAS dont les logiciels de gestion sont étroitement liés au matériel sous-jacent.

En parallèle nous avons étudié des solutions purement logicielles. Nous souhaitions pouvoir installer ces briques logicielles sur des serveurs standards et disposer d'une interface conviviale et unifiée pour administrer les hyperviseurs, le stockage et les *VM*.

2.1.1 Le couplage vSphere/vSAN

vSphere, que nous connaissons bien, s'impose en tant qu'hyperviseur. En effet, quand un environnement de virtualisation est prérequis par un éditeur, c'est presque toujours vSphere.

vSAN[14] (pendant « virtualisation du stockage » entièrement intégré à vSphere), était de plus en plus déployé et son déploiement était facilité depuis sa version 6.5. Avec VxRail, DELL/EMC a fait le choix d'intégrer vSAN et vSphere. Dans la communauté ESR les premiers retours semblaient satisfaisants [1] [2], ce qui achevait de nous convaincre de la possibilité de le déployer nous-mêmes, et ainsi de disposer de support pour la virtualisation et le stockage par le même éditeur bien connu.

vSAN implémente le codage à effacement (*erasure coding*), ce qui impose de n'utiliser que des SSD. C'est une méthode de protection des données dans laquelle celles-ci sont divisées en *k* fragments, développés et codés avec *m* éléments de données redondants et stockés sur un ensemble de différents emplacements ou supports de stockage. Cette technique qui associe une parité aux données, ressemble à du RAID5 et permet d'économiser de l'espace tout en assurant de la redondance. Il est possible de perdre jusqu'à *m* fragments de données, leur contenu est alors décodé par le code d'effacement (cf. Figure 1).

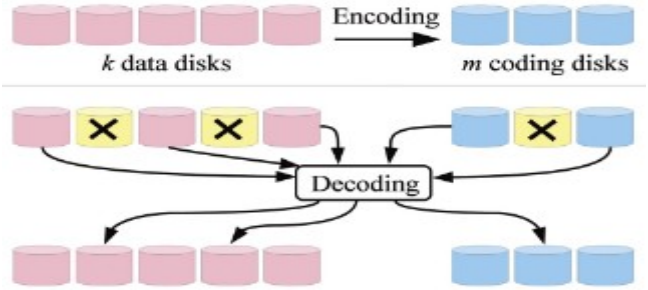

Figure 1: fonctionnement de l'erasure coding [3]

Toutefois, vSAN reste limité en fournissant uniquement du stockage pour les *VM* (le mode fichier n'étant pas assez étoffé et le mode stockage objet inexistant), et après une formation de l'équipe Système, nous avons déployé le couple vSphere/vSAN dans une infrastructure hyper-convergée.

2.1.2 Le stockage à définition logicielle avec Ceph

La solution de stockage distribué libre Ceph a déjà été présentée à plusieurs reprises aux JRES [4] [5] [6] [7] [8] [9] et lors du tutoJRES n°18 dédié au stockage distribué [10] [11].

Ceph s'installe sur du matériel standard (*commodity hardware*) sous Linux, son architecture distribuée facilite le passage à l'échelle par l'ajout de disques pour étendre le stockage, ou par l'ajout de nœuds et de contrôleurs disques, pour en augmenter conjointement les performances. Le logiciel libre Ceph et sa communauté adressent les manques d'un SAN, en termes d'élasticité et de souplesse et il dispose de solides références (RedHat, Suse, le CERN, de nombreuses universités, laboratoires, Orange, Crédit Mutuel…).

vSphere ne dispose pas du support Ceph natif. Nous pensions alors intégrer le stockage Ceph aux hyperviseurs vSphere, à travers iSCSI ou NFS, pour servir les disques de nos *VM*, mais cela nécessitait l'empilement de couches logicielles : iSCSI+RBD ou NFS+CephFS.

14 Solution de stockage distribué de VMware

SUSE Enterprise Storage [12] (produit de stockage construit sur Ceph) était alors seul à proposer le déploiement d'une passerelle iSCSI [13] au-dessus de RBD (désormais supporté par RedHat Ceph Storage [14]) et nous doutions des bonnes performances de l'intégration au niveau des entrées/sorties. La Gateway iSCSI de Ceph présente une cible iSCSI hautement disponible qui exporte les images RBD en tant que disques SCSI (sans support de *SCSI persistent reservation* [15] pour contrôler l'accès à des disques partagés) vers les clients vSphere.

Par ailleurs, nous avons exclu d'exporter CephFS vers NFS qui aurait induit une forte latence et aurait constitué un goulot d'étranglement critique pour les entrées/sorties des *VM*.

Aussi, faute de pouvoir intégrer Ceph dans vSphere pour faire tourner nos *VM*, nous nous sommes intéressés aux technologies de virtualisation opensource qui auraient pu tirer parti de Ceph nativement. Nous avons choisi l'hyperviseur KVM (solution de virtualisation libre), déjà bien connu et largement déployé dans des structures de toute taille.

2.1.3 L'hyperviseur KVM

KVM est un hyperviseur libre intégré au noyau Linux depuis 2007. Couplé à QEMU[15], il est très répandu, éprouvé et performant. Il peut faire tourner des *VM* sous Windows, Linux ou *BSD et l'utilisation de pilotes *virtio* dans la *VM* accélère leur exécution. Très bien interfacé avec Ceph, QEMU permet l'utilisation d'images de périphériques de blocs directement sur des pools Ceph.

OpenStack, OpenNebula, RedHat Virtualization font partie des nombreuses infrastructures de virtualisation qui s'appuient sur KVM depuis plus de 10 ans.

2.1.4 Proxmox Virtual Environment pour administrer Ceph et KVM

Proxmox Virtual Environment est une solution de virtualisation libre basée sur l'hyperviseur Linux KVM, construite sur Debian depuis 2008. Elle propose un support payant, auquel nous avons souscrit, pour accéder aux dépôts *Enterprise* (cycle de validation plus long) et à différents niveaux d'intervention. Notons toutefois que le support Proxmox n'inclut pas la récupération de données.

Dans Proxmox se retrouvent presque toutes les fonctionnalités présentes dans vSphere (mécanismes de *HA*, migration à chaud).

DRS[16] n'est pas implémenté dans Proxmox, mais, avec son API riche et l'accès à toutes les fonctionnalités sans licences additionnelles, il va plus loin que VMware (cf.Figure 6) :

– possibilité d'accéder à de nombreux *back-ends* de stockage pour les *VM* (NFS, SMB/CIFS, GlusterFS, RADOS[17] Block Device[18] (RBD), CephFS, ZFS, iSCSI) ;

– chaque nœud permet de gérer le cluster, contrairement à vSphere où il faut déployer une *VM* dédiée à ce rôle et acquérir la licence adéquate ;

– gestion des *VM* facilitée, la configuration de chacune étant stockée dans un fichier plat (qui contient également les modifications de l'enveloppe de la *VM* faites dans un *snapshot*) et est plus simple qu'avec vSphere ;

– intégration native de la possibilité de sauvegarder des *VM* ;

15 Logiciel libre de machine virtuelle pouvant émuler un processeur ou, à l'aide de KVM, virtualiser un ou plusieurs systèmes d'exploitation

16 Mécanisme spécifique à vSphere qui permet notamment de répartir dynamiquement les VM sur les différents hôtes en fonction de leurs charges

17 Ceph RADOS (*Reliable Autonomic Distributed Object Store*) : moteur de stockage objet fiable autonome et distribué

18 Interface de stockage en mode bloc qui agrège les données réparties sur plusieurs disques (OSD) dans un cluster Ceph et permet entre autres la création de snapshots, la réplication, l'export

– disponibilité d'un pare-feu basé sur *iptables*[19] et activable pour chaque *VM* (ainsi qu'au niveau des hôtes et du cluster) ;

– enfin, très bonne intégration de Ceph, de nombreuses opérations sur Ceph peuvent s'exécuter directement depuis l'interface graphique. Précisions toutefois que Proxmox n'implémente pas nativement l'*erasure coding* dans son interface graphique.

Proxmox semble de plus en plus populaire dans la communauté, avec un premier retour d'expérience en 2017 [16], mais également chez des hébergeurs privés.

3 Choix d'une double *HCI* Proxmox/Ceph + vSphere/vSAN

Comme vSphere s'impose pour répondre à certaines exigences d'éditeurs, mais que Ceph+Proxmox répondent mieux à l'élasticité dont nous avons besoin, nous avons choisi de ne pas mettre tous nos nœuds dans le même panier et de déployer ces deux *HCI* concurrentes en parallèle.

Nous avons installé ces *HCI* sur un socle de matériel identique, afin de pouvoir se donner la possibilité de redéployer des hyperviseurs ou des SSD dans l'une ou l'autre des infrastructures le cas échéant. De plus cela nous permet de comparer les deux en conditions réelles.

Comme toujours en pareilles circonstances, au fur et à mesure que nous avançons dans nos études, des questions surgissent : « Ceph sera-t-il facile à prendre en main ? KVM sera-t-il fiable ? Proxmox sera-t-il stable ? Arriverons-nous à migrer nos *VM* de vSphere vers KVM ? Serons-nous capables de migrer des *VM* de KVM vers vSAN ? Le support sera-t-il à la hauteur ? Serons-nous capables de prendre en main ces deux *HCI* ? Enfin ces *HCI* répondront-elles aux besoins croissants des projets portés par la DSI ? »

3.1 Trois centres de données et des équipements réseau en commun

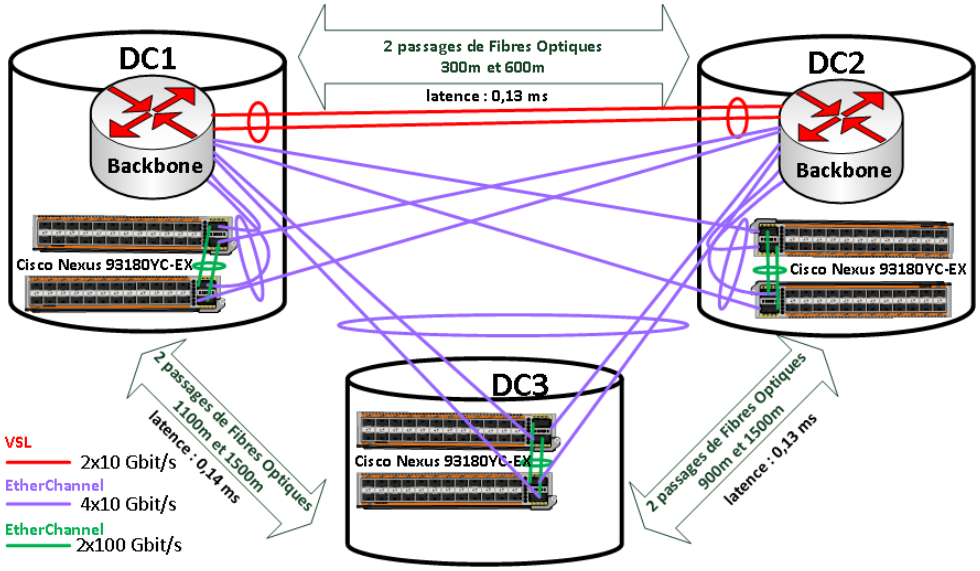

Figure 2 : Schéma de l'architecture réseau de la double *HCI*

Les deux infrastructures hyper-convergées ne sont pas totalement disjointes : elles sont localisées dans les trois mêmes centres de données qui les hébergent (DC1, DC2 et DC3). Elles partagent les mêmes équipements réseau et les mêmes fibres optiques, ce qui peut être vu comme un SPOF[20] de la double *HCI* (cf. Figure 2), ou de la mutualisation, considérant l'architecture réseau résiliente.

19 Interface en ligne de commande permettant de configurer les règles de *netfilter*, le pare-feu sous Linux

20 Point unique de défaillance : point informatique dont e reste du système dépend, et dont une panne entraîne l'arrêt complet du système

Le déploiement de la première infrastructure s'est effectué progressivement sur un, puis deux puis trois sites, alors que la seconde a été déployée directement sur les trois sites.

Nous avons d'abord choisi de déployer deux commutateurs de 48 ports en 10/25 Gbit/s (Cisco Nexus 93180YC-EX [17]) en mode vPC[21] dans chaque salle, pour irriguer les différents nœuds.

Les liens entre les deux *switches* au sein d'un même centre de données sont en 2 × 100 Gbit/s.
Les interconnexions entre les commutateurs et le cœur de réseau (*backbone*) sont en 4 × 10 Gbit/s.
Les interconnexions entre les deux *switches* qui constituent le *backbone* sont en 2 × 10 Gbit/s.

La latence mesurée représente la moyenne de la latence de chaque LACP.
Le DC1 est raccordé au DC2 par 2 fibres noires de 300 et 600 m, avec une latence de 0,13 ms.
Le DC2 est raccordé au DC3 par 2 fibres noires de 900 et 1 500 m, avec une latence de 0,13 ms.
Le DC1 est raccordé au DC3 par 2 fibres noires de 1 100 et 1 500 m, avec une latence de 0,14 ms.

C'est parce que l'université disposait d'un cœur de réseau performant, résilient et maîtrisé que nous avons pu envisager de déployer des solutions de stockage distribuées.

3.2 Unique socle de matériels homogènes

La spécification matérielle des serveurs a été compliquée par VMware qui impose de fortes contraintes sur les composants pouvant faire tourner vSAN. Ils doivent être validés dans la matrice de compatibilité VMware [18], contrairement au couple Proxmox/Ceph qui s'installe partout.

Nous avons fait le choix d'un stockage uniquement sur SSD Read Intensive (supportant jusqu'à une écriture complète par jour pendant cinq ans) disposant d'une interface SAS[22].

Nous souhaitions des CPU avec une fréquence élevée dont Ceph et vSAN pourraient tirer parti et d'un grand nombre de cœurs pour la virtualisation et des cartes réseaux 10 Gbit/s performantes pour optimiser le stockage. Nous avons choisi une carte RAID simple, qui permet néanmoins de présenter les disques en JBOD[23], dont nous aurions pu nous dispenser pour une simple carte HBA[24].

Le fonctionnement de vSAN par *diskgroups* (1 disque de cache pour 5 à 7 disques de capacité) a imposé l'achat de disques SSD de 800 Go *Write Intensive* pour le cache (supportant jusqu'à 10 écritures complètes par jour pendant 5 ans).

L'agrégation des liens n'est pas imposée par les logiciels (qui assurent eux-mêmes la redondance), mais nous souhaitions pouvoir agréger tous les ports réseau deux à deux avec le protocole LACP[25], afin de disposer d'une plus grande résilience des clusters et pouvoir absorber des pics de charge. Sur vSphere/vSAN un agrégat est dédié à vMotion[26] et au réseau d'administration, un autre pour vSAN et un troisième aux réseaux des *VM*. Sur Proxmox/Ceph, un agrégat est dédié à Proxmox, un second au réseau public de Ceph, un troisième au réseau du cluster Ceph et un quatrième aux réseaux des *VM*. En effet, Ceph préconise d'utiliser 2 réseaux distincts : 1 pour le stockage et l'autre pour le cluster Ceph, alors que vSAN n'en nécessite qu'un seul, ce qui a impliqué l'acquisition de 2 ports réseaux supplémentaires pour chaque machine (soit 2 cartes 4 ports par serveur).

Ce plus grand dénominateur commun a augmenté le coût unitaire de chaque serveur, mais ces coûts sont mesurés et c'était le prix à payer pour pouvoir basculer vers l'une ou l'autre des infrastructures.

21 Technologie d'agrégation d'interfaces réseau sur plusieurs commutateurs, disponible sur les Cisco Nexus

22 Serial Attached SCSI : interface disque dont le contrôleur peut envoyer 2 commandes en même temps (duplex), et propose les meilleurs débits

23 Just a Bunch Of Disks : disques présentés tels quels au système, sans activer de mécanismes RAID

24 Host Bus Adapter : le contrôleur hôte de bus est une carte d'extension qui permet ici de connecter des disques durs

25 Technique d'agrégation de liens pour regrouper plusieurs ports réseau et les utiliser comme s'il s'agissait d'un seul (référence IEEE 802.3ad)

26 Dans un cluster vSphere, c'est la possibilité de migrer à chaud une VM (en exécution) d'un hôte à un autre. Proxmox nomme cela live migration

Nous souhaitions investir progressivement et étendre la capacité petit à petit du cluster, afin de gérer la fin de vie et la rotation des matériels. Cependant le manque d'investissements dans les infrastructures de la DSI depuis de nombreuses années poussait l'arrivée de plus gros budgets précocement qui nous ont contraints à déployer plus vite un grand nombre de nœuds. Cela nous amènera toutefois à anticiper un renouvellement massif (cf. Figure 7).

3.3 Quelques divergences matérielles

3.3.1 Trois moniteurs Ceph dédiés

Afin de rendre plus robuste l'infrastructure Ceph en respectant les bonnes pratiques, nous avons déployé un moniteur Ceph dédié dans chaque centre de données. En effet, en cas de *crash* et de reconstruction du stockage Ceph, ce service est très sollicité au même titre que les OSD[27], il est donc conseillé de dissocier les serveurs physiques. Ce petit serveur est intégré au cluster Proxmox comme les autres nœuds, mais n'exécute pas d'OSD et ne peut pas héberger de *VM* (cf. Figure 8).

3.3.2 Un *Witness* pour vSphere

La *HCI* de VMware fonctionne sur trois sites, mais seuls deux sites sont utilisés pour le stockage ou la virtualisation des *VM*. Le troisième site héberge un serveur *witness* (cf. Figure 9). Dans le *stretched cluster*[28]vSAN que nous avons déployé, c'est l'hôte témoin qui contient les métadonnées des *VM*. Le *witness* stocke les composants témoins de chaque objet de *VM*, et en cas de panne d'un site, il pourra former un quorum avec le site vivant, évitant ainsi le problème de *split-brain*[29].

3.4 L'infrastructure Proxmox/Ceph

Après avoir défini les matériels et les logiciels souhaités, nous avons lancé les premiers investissements. Les clusters Proxmox/Ceph ont formé la première infrastructure que nous avons déployée progressivement, car c'est la première technologie de stockage distribué sur laquelle nous nous sommes formés. C'est aussi celle dont le saut technologique était le plus important, qui levait le plus d'interrogations et vis-à-vis de laquelle nous avions le plus d'attentes.

3.4.1 Un petit cluster Proxmox/Ceph de sauvegarde devenu le socle du PRA/PCA des *VM* Proxmox

Dès 2017, nous avons déployé un premier cluster Ceph sur quatre serveurs physiques (cf. Figure 10) composés de disques mécaniques capacitifs, qui constituent aujourd'hui le cluster de sauvegarde et le socle technique du PRA[30]/PCA[31] des *VM* de type KVM. Des *snapshots* RBD de chaque image disque des *VM* en production sont pris tous les jours puis exportés et historisés vers ce cluster externe à la production. Les 5 derniers *snapshots* sont conservés sur le site de production, sans être accessibles aux utilisateurs depuis l'interface Proxmox.

Pour pouvoir déployer un PRA ou un PCA (et se protéger en cas de destruction totale du cluster Ceph), nous exportons également les fichiers de configurations des *VM* sur chaque nœud du cluster de sauvegarde, ainsi en quelques minutes nous pouvons redéployer une *VM* dans un environnement Proxmox/Ceph fonctionnel, mais dont les performances sont dégradées. Pour l'instant la procédure de reprise en cas de sinistre sur Ceph est manuelle et le script d'automatisation en chantier.

27 Dans Ceph chaque disque de stockage est associé à un démon Object Storage Daemon (OSD)

28 Un cluster étendu est un cluster vSAN déployé sur 2 sites. Les 2 sites sont actifs et répliqués entre eux. Si un site est indisponible vSAN utilise le stockage sur l'autre site et vSphere HA redémarre les VM concernées sur le site actif.

29 Un split-brain se produit lorsque 2 parties d'un cluster dont déconnectées, chaque partie croyant que l'autre ne fonctionne plus.

30 Plan de Reprise d'Activité : ensemble de procédures qui permettent de prévoir la reconstruction d'un système d'information en cas de sinistre

31 Plan de Continuité d'Activité : procédures pour redémarrer le plus rapidement possible avec le minimum de pertes de données en cas de sinistre

3.4.2 L'architecture Proxmox/Ceph de production

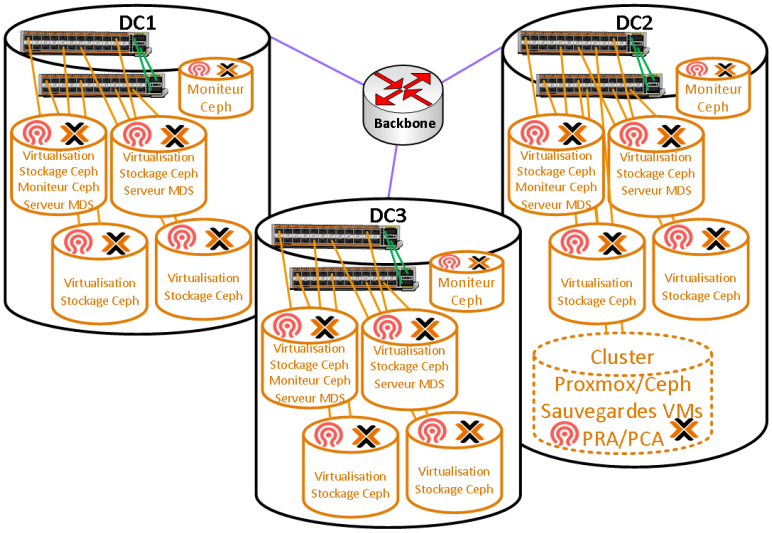

Figure 3 : Schéma d'architecture de l'infrastructure Proxmox/Ceph

3.4.3 Phasage du déploiement de l'*infrastructure Proxmox/Ceph*

En 2019 nous avons créé le cluster Proxmox/Ceph de production sur quatre nœuds localisés dans un centre de données, avec migration et déploiement des premières *VM* pour alléger l'ancien cluster vSphere stocké sur Compellent et dont les ressources étaient saturées. En 2020 nous avons ajouté quatre nœuds dans chacun des deux autres centres de données, ainsi qu'un moniteur Ceph dédié sur chaque site. En 2021 nous avons également ajouté des SSD dans chaque nœud.

3.4.4 Rôles des nœuds Proxmox et Ceph

Dans le cluster Proxmox, tous les nœuds ont le même rôle de moniteur et participent de façon égale au quorum. Synchronisés, ils servent tous également d'interface de gestion du cluster.

Le mécanisme *HA* a été activé sur tous les nœuds sauf sur les trois moniteurs Ceph dédiés qui ne virtualisent pas de serveurs. Ils pourraient le faire, mais nous n'y avons pas configuré les réseaux des *VM* et ne les avons pas intégrés au groupe *HA* afin d'éviter ce cas. La migration à chaud est donc rendue impossible et en cas de migration à froid, le mécanisme *HA* redémarrerait automatiquement les *VM* concernées sur un nœud fonctionnel.

De plus certains nœuds Proxmox/Ceph jouent des rôles particuliers au sein du cluster. Ceph préconise la configuration de trois moniteurs minimum et jusqu'à cinq pour les gros clusters (de façon à assurer un quorum), en privilégiant les moniteurs dédiés. C'est pourquoi nous avons choisi de configurer deux autres nœuds mixtes avec le rôle annexe de moniteur Ceph.

Afin d'héberger le service Moodle[32] en cluster, nous avons également déployé le système de fichiers distribués CephFS[33]. Il s'appuie sur un ou plusieurs services de métadonnées (MDS[34]), que nous avons déployés sur six nœuds en mode Actif/Passif (deux dans chaque centre de données).

Nous avons fait le choix de donner une priorité plus faible en cas de *HA* aux nœuds de stockage configurés en moniteurs et/ou serveurs de métadonnées.

32 Plateforme libre d'apprentissage en ligne utilisée par les étudiants de l'université

33 Système de fichiers Ceph compatible POSIX construit sur RADOS et qui peut être partagé entre des serveurs

34 Démon serveur de métadonnées nécessaire au fonctionnement de CephFS

3.4.5 Options de configuration du stockage Ceph

Nous n'avons pas opté pour l'*erasure coding* qui nous aurait permis d'économiser de l'espace de stockage, car ce n'était pas supporté par Proxmox avant la version 7.2 (mai 2022). Nous utilisons des pools répliqués trois fois (comme préconisé par Ceph), qui consomment donc 300 % de l'espace nécessaire, mais nous pourrions appliquer également d'autres profils de résilience moins gourmands (deux réplicas voire un seul) pour des données jetables par exemple.

L'algorithme de placement CRUSH[35] a été configuré avec un domaine de panne qui assure la localisation d'un réplica dans chacun des trois centres de données. De plus, en cas d'isolation réseau, de sinistre ou de panne d'un site, les réplicas se répartissent sur des nœuds et des disques différents. Cette politique assure la disponibilité en écriture du stockage en cas de perte d'1/3 du stockage (1 site, 4 nœuds ou 76 SSD) et de la disponibilité en lecture et en possible reconstruction en cas de perte de 2/3 du stockage (2 sites, 8 nœuds ou 152 SSD). **Tant qu'il reste un seul site vivant (ou 4 nœuds) aucune donnée n'est perdue**. Il convient toutefois de valider la bonne répartition des PG[36] conformément à la configuration des règles CRUSH.

3.4.6 Capacité du cluster Proxmox/Ceph

Un cluster Ceph nécessite un minimum de 3 nœuds. En répliquant les données dans 3 centres de données, il faut au moins 9 nœuds pour se protéger de la perte de 2/3 de l'infrastructure.

En 2022, le cluster compte 4 nœuds hyper-convergés de virtualisation et de stockage ainsi qu'un nœud moniteur Ceph dédié dans chaque centre de données. Soit pour les 12 nœuds répartis sur les 3 sites, un total de : 864 vCPU, 9 To de vRAM pour un stockage brut de 480 To (228 OSD de 1,9 To).

Actuellement 260 *VM* tournent sur l'ensemble de cette infrastructure. Aussi, en réservant 20 % de ressources pour Ceph et Proxmox, et avec notre politique de stockage, elles disposent environ de 700 vCPU, 7 To de vRAM et 160 To de stockage disponible.

3.4.7 Évolution des versions de Ceph et Proxmox au fil du temps

Depuis le début du déploiement, nous avons mis à jour Ceph à plusieurs reprises dans ses versions majeures, avec une installation initiale sous Luminous (12.2), puis passage à Mimic (13.02), Nautilus (14.02), Octopus (15.02) et enfin Pacific (16.02). Proxmox, initialement déployé en version 5.4 a également été mis à jour vers la version 6, puis 7.1.

Toutes ces mises à jour n'ont nécessité aucun arrêt de service. La documentation de Proxmox, adaptée de celle de Ceph, a toujours été simple, fiable et efficace pour effectuer chaque mise à jour.

3.5 L'infrastructure vSphere/vSAN

Lors du déploiement du cluster vSAN/vSphere, nous avons établi des règles objectives de répartition des *VM* sur l'une ou l'autre des infrastructures, afin de pouvoir déterminer sur quelle infrastructure tourne une *VM*.

3.5.1 Rôle des nœuds vSphere/vSAN

À l'exception du serveur *witness*, tous les nœuds du cluster VMware jouent le même rôle (stockage + virtualisation). L'administration passe par le déploiement d'un serveur vCenter[37] et l'acquisition de la licence d'utilisation, que nous avons déployé en mode *HA* entre DC1 et DC3.

35 *Controlled, Scalable, Decentralized Placement of Replicated Data*, c'est un algorithme de placement des données pseudo aléatoire qui distribue les données de façon efficiente et robuste au sein du cluster structuré [19] .

36 Dans Ceph un Groupe de Placement est une collection logique d'objets répliqués sur les OSD pour assurer la fiabilité du stockage

37 vCenter Server Appliance (vCSA) est une VM préconfigurée et optimisée pour l'exécution de vCenter et des services associés

3.5.2 L'architecture vSphere/vSAN

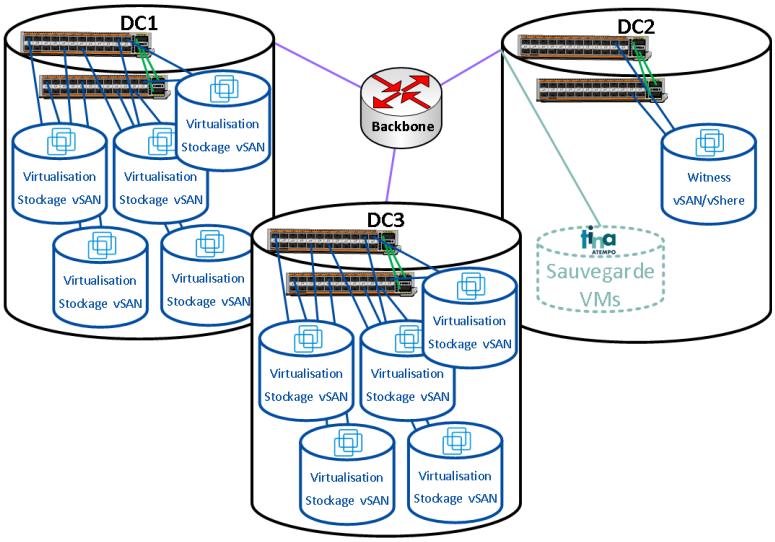

Figure 4 : Schéma d'architecture de l'infrastructure vSphere/vSAN

3.5.3 Phasage du déploiement vSphere/vSAN

Le cluster vSphere/vSAN a été créé en 2020 sur deux centres de données comptant quatre nœuds chacun, avec le déploiement de l'hôte *witness* sur le troisième site (minimum pour tolérer une panne sur un site). Les premières *VM* y ont été migrées fin 2020. En 2021, nous avons ajouté un nœud de stockage et de virtualisation sur les deux sites de production, ainsi que des SSD. Nous avons choisi de déployer directement vSphere et vSAN en version 7.0 publiée dans l'année, afin d'économiser une mise à jour majeure ultérieurement. Nous avons essuyé quelques bugs liés à la publication précoce, ce qui nous a conduits par la suite à reprendre nos habitudes de laisser passer un temps long après la sortie d'une nouvelle version de vSphere.

3.5.4 Options de configuration du stockage vSAN

Pour fonctionner, vSAN construit un magasin de données (*datastore*) unique entre tous les nœuds du cluster à partir de groupes de disques (*diskgroup*). Chaque nœud est composé de 3 *diskgroups* eux-mêmes construits avec 5 SSD pour le stockage et 1 SSD de cache (800 Go).

vSAN est configuré en *stretched cluster* : les deux sites sont en miroir, avec une réplication des données sur chaque site. Nous avons opté pour une politique de stockage utilisant l'*erasure coding* en mode RAID5/FTT1, ce qui permet de tolérer la perte d'un seul tronçon de données sur un site.

Cette politique assure la disponibilité en écriture du stockage en cas de panne d'un site de production (DC1 et DC3) ainsi que d'un serveur sur le site restant. En revanche, le stockage serait perdu en cas de sinistre sur les deux sites de production, contrairement à Ceph.

Cette politique de stockage consomme 266 % de l'espace nécessaire.

3.5.5 Capacité du cluster vSphere/vSAN

En 2022, le cluster compte 5 nœuds hyper-convergés de virtualisation et de stockage répliqués sur 2 centres de données + 1 nœud *witness* sur un 3e site (DC2). Soit pour les 10 nœuds répartis sur les 2 sites, un total de : 720 vCPU, 7,5 To de vRAM pour un stockage brut de 260 To (30 *diskgroups* de 8 To). Actuellement 290 *VM* tournent sur l'ensemble de cette infrastructure. Aussi en réservant 20 % de ressources pour vSAN et vSphere, et avec notre politique de stockage, elles disposent environ de 570 vCPU, 6 To de vRAM et 100 To de stockage disponible.

3.5.6 Sauvegarde des *VM* sous vSphere

Nous sauvegardons les *VM* avec le logiciel TiNa HyperVision Deduplication Storage de Atempo qui est connecté au serveur vCenter. Actuellement le serveur physique hébergeant l'application ADE-HVDS s'appuie sur une baie de disques (comme nos actuels serveurs de sauvegarde), et nous avons validé la réplication de ces données sur un petit cluster Ceph indépendant. Ce mode opératoire est officiellement supporté par Atempo, en utilisant le mode bloc de Ceph (RBD). De plus, l'éditeur nous a accompagnés pour sauvegarder les fichiers de CephFS avec TiNa.

4 Joies et désenchantements de l'exploitation de deux *HCI*

Les infrastructures étant indépendantes, il a fallu nous former à de nouvelles technologies de stockage différentes et cela a retardé leurs mises en service. Toutefois la connaissance d'une technologie permet d'appréhender beaucoup plus facilement la seconde. Cette maîtrise technologique était un enjeu majeur dans notre projet, afin d'être en mesure de répondre aux demandes croissantes d'hébergement de services sécurisés.

Les études ont été faites en rencontrant des constructeurs, des éditeurs, des intégrateurs, des administrateurs, en sondant les listes métiers et en épluchant les retours d'expérience. Nous avons suivi des formations sur vSphere, le stockage distribué, Ceph, Proxmox et vSAN.

Nous avons automatisé la migration des *VM* de l'ancien cluster vers les deux nouvelles infrastructures, qui ont ainsi pu être redéployées en trois mois, à l'aide de PowerCLI de vSphere et de l'API Proxmox.

Nous évaluons la quantité de travail à 400 jours/hommes sur 4 ans pour les études préalables, la formation, la spécification matérielle, l'acquisition, le déploiement des 26 serveurs et la migration des 500+ *VM* de la double architecture. Ce projet a impliqué 4 personnes, une pour chaque sujet : Proxmox/Ceph, vSphere/vSAN, Réseau, *VM*.

L'exploitation et le MCO des deux *HCI* est comparable aux autres solutions, et l'équipe qui a reçu formation et transfert de compétences sur les technologies déployées est compétente pour opérer les deux architectures, avec l'aide du support le cas échéant. La pile logicielle Ceph consomme plus de temps, car il propose plus de services associés en plus de la fourniture de stockage pour les *VM* (système de fichiers, stockage objet, passerelle iSCSI) dont nous souhaitons profiter.

4.1 Les avantages d'opérer deux *HCI*

L'exploitation conjointe de deux infrastructures d'hébergement de *VM* nous semble un atout pour augmenter la disponibilité et la résilience des services hébergés en clusters : en effet les différents nœuds du cluster applicatif peuvent fonctionner en parallèle (actif/actif) sur les deux infrastructures et le service n'est pas impacté en cas d'indisponibilité d'une *HCI* (cf. Figure 5).

Par ailleurs, dès le début du projet nous avons envisagé la possibilité de pouvoir déménager les *VM* de vSphere vers KVM et réciproquement, notamment pour le cas où l'on ne conserverait qu'une seule solution. L'opération de déplacement est scriptée de vSphere vers Proxmox (ajout des pilotes virtuels *virtio*, conversion des disques avec *qemu-img*, reconfiguration des interfaces réseau, installation du nouvel agent) et il faut 10 minutes pour migrer une *VM* de 50 Go. C'est encore en cours de mise en œuvre de Proxmox vers vSphere.

Cela nous donne la liberté de déployer des *VM* sur l'infrastructure la plus appropriée en fonction de la matrice de compatibilité du service déployé, du coût associé (licences), des performances et de la stabilité.

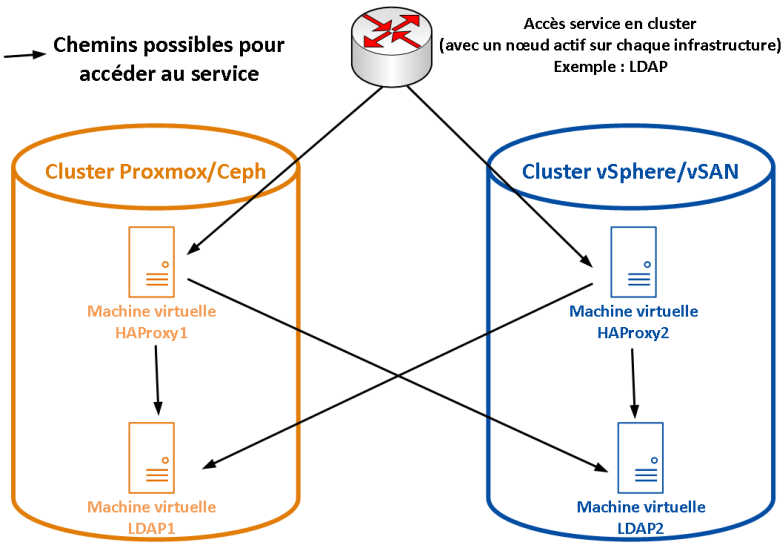

Figure 5 : Schéma de l'architecture de la double HCI Proxmox/Ceph + vSphere/vSAN

La création des systèmes Linux automatisée par PXE[38] est indépendante de l'infrastructure de virtualisation et le système est instantanément disponible sur chaque infrastructure. Seuls les systèmes Windows sont déployés à partir d'un *template* porté par une infrastructure et migrés sur l'autre le cas échéant.

L'administration des Systèmes dans les *VM* invitées se fait de la même façon avec des outils communs : pour la gestion des configurations, la sauvegarde des fichiers, la journalisation centralisée, la supervision, la métrologie…

Nous avons également la possibilité de redéployer des disques d'une infrastructure à l'autre, au cas où l'une nécessiterait plus de stockage que l'autre, ou pour absorber un déséquilibre.

Ceph est plutôt orienté résilience du stockage, cependant les tests effectués avec *fio*[39] montrent qu'il est plus performant que vSAN, sur les opérations de lectures/écritures aléatoires, notamment sur les opérations d'écritures (cf.Figure 11). Les tests ont été réalisés avec la même *VM* sous RockyLinux (2 vCPU et 4 GO de RAM) d'abord sur l'une puis déplacée sur l'autre des *HCI*.

4.2 Et les inconvénients

La principale limite de notre double infrastructure d'hébergement est qu'elle est construite sur le même socle réseau. Une opération de maintenance du cœur de réseau a rendu les deux *HCI* inopérantes pendant quelques minutes. Les stockages ont néanmoins été reconstruits automatiquement au retour de la connexion. À ce jour nous n'envisageons pas d'évolution pour combler cette faiblesse.

Deux technologies différentes impliquent au moins deux interfaces de gestions différentes, deux documentations d'exploitation à maintenir, ce qui nécessite une courbe d'apprentissage plus longue, qui a été absorbée par les nouveaux collègues récemment intégrés à l'équipe.

La répartition des clusters applicatifs au-dessus des deux *HCI* doit être bien validée et peut amener des difficultés lors de pannes non franches où la *VM* semble vivante (réponse au *ping*, port applicatif ouvert…) et reçoit des requêtes sans rendre le service.

38 L'environnement d'exécution *pre-boot* permet à un système de démarrer depuis le réseau

39 Outil simulant une charge de travail pour mesurer les performances du stockage sous Linux.

4.3 Quelques déconvenues avec vSphere/vSAN

Le coût des multiples licences VMware (vSphere + vSAN + vCenter + éventuellement NSX[40]) et du support élevé rend cette infrastructure moins éligible à être étendue ni reproduite sur de plus petits clusters, au profit de Proxmox/Ceph dont les coûts logiciels peuvent être nuls en utilisant les dépôts communautaires de Proxmox (*No-Subscription Repository*) et sans contracter de support.

Comme nous l'avons vu, le cluster VMware s'administre via vCenter - une unique *VM* dédiée - ce qui fait peser des contraintes de disponibilité sur cet *appliance*, que l'on peut résoudre en activant *vCenter HA*.

La perte d'un SSD de cache sur l'infrastructure VMware entraîne la reconstruction coûteuse en ressource de tous les objets du *diskgroup*. Alors que dans Ceph tous les OSD (construits sur des SSD) sont identiques et la perte d'un SSD n'entraîne que la reconstruction des données hébergées sur l'OSD concerné. Avec vSAN, le mécanisme de *diskgroup* nécessitant un disque de cache (1 *diskgroup* = 1 cache + 5 à 7 disques) est moins souple que Ceph (1 OSD = 1 disque) qui permet l'ajout de disques de façon unitaire.

4.3.1 Retour sur un incident vSphere

Afin d'anticiper une mise à jour ultérieure, nous avons récupéré trop rapidement après sa sortie la version 7.0 *Update 2* de vSphere : VMware l'a supprimée des dépôts au bout d'une semaine, car elle avait un « problème ». Deux mois plus tard, quand nous avons appliqué cette mise à jour, nous nous sommes retrouvés bloqués. L'indisponibilité de cette image sur les dépôts VMware invalidait un certain nombre d'actions critiques dans l'interface vSphere : notamment la prise de snapshots, la mise en maintenance ou le mécanisme de *HA*. Conjointement, nous avons observé une panne de certaines *VM* qui n'ont pas pu redémarrer automatiquement car HA n'était pas fonctionnel. D'après VMware, il semble qu'une micro-coupure du serveur *Witness* en soit la cause (alors qu'il ne fait qu'arbitrer entre les deux salles de production et peut normalement être arrêté à tout moment).

Cette panne a fait apparaître un dysfonctionnement dans l'organisation du support VMware. N'arrivant ni à réactiver *HA*, ni à mettre un nœud en maintenance, ni à effectuer de mise à jour, nous avons ouvert un ticket auprès du support VMware qui a lui-même ouvert plusieurs tickets, sans gestion centralisée, auprès d'autres services supports VMware : vSphere, vSAN, vCenter. En effet, chaque couche logicielle VMware est supportée indépendamment, alors que nous pensions le contraire en prenant la pile complète, virtualisation et stockage, chez le même éditeur.

4.4 Un atout pour Proxmox et un pour Ceph

Dans le cas des *VM* hébergées pour des laboratoires ou des services extérieurs à la DSI, il est indispensable de pouvoir filtrer les flux des *VM* au sein d'un même sous-réseau qui ne sont pas protégés par les pare-feu du cœur de réseau.

Aussi pour ces *VM* nous ciblons systématiquement l'infrastructure Proxmox/Ceph, car à ce jour c'est la seule infrastructure où nous avons déployé un pare-feu sur les *VM* depuis le cluster de virtualisation. Les *VM* hébergées sont ainsi isolées entre elles au sein d'un même VLAN[41]. En effet, nous avons choisi de ne pas déployer NSX à cause du coût des licences élevées.

Proxmox permet facilement l'intégration d'un cluster Ceph externe et il est aisé d'exporter et d'importer des disques de *VM* (images RBD) ou des pools complets d'un cluster à l'autre.

40 Plateforme VMware de virtualisation du réseau permettant notamment de servir de pare-feu entre les VM

41 Réseau local virtuel

4.4.1 Retour sur un incident Proxmox

Lors du déploiement d'un moniteur Ceph (nœud intégré au cluster Proxmox), sur un nouveau *switch*, la MTU[42] avait été laissée à sa valeur d'origine et non positionnée à 9000 comme attendu sur le VLAN dédié à Proxmox.

Ce nœud a provoqué des dysfonctionnements à cause des pertes de messages de *corosync*[43] utilisé par le cluster Proxmox pour construire un quorum.

Cela conduisait à des redémarrages intempestifs de certains nœuds et générait le déclenchement de *HA* avec le redémarrage des *VM* qui étaient hébergées sur le nœud fraîchement redémarré. La plupart des hyperviseurs ont ainsi redémarré deux à trois fois pendant quelques heures avant de se stabiliser.

Le support de Proxmox, réactif, maîtrisait l'ensemble des couches Proxmox/Ceph et nous a permis de diagnostiquer rapidement le problème pour corriger l'erreur de configuration.

4.4.2 Retour sur un incident Ceph

Lors de l'extension du cluster Ceph sur les trois salles, la règle de placement des PG dans le cluster n'était pas correctement prise en compte par le cluster et des PG se trouvaient localisés sur une seule salle [20]. Aussi lors d'une opération de déplacement de disques, 4 PG (sur 6689 en tout) ont été temporairement inaccessibles, rendant indisponibles les pools concernés (métadonnées de CephFS et pools de *VM*).

Le support Proxmox n'étant pas qualifié pour la récupération de données (qui pourtant n'étaient pas réellement perdues), nous avons dû faire appel à un prestataire Ceph pour nous accompagner.

La suppression des OSD dans Ceph (« *ceph osd purge* ») n'est pas destructrice de données, nous avons donc pu récupérer les PG concernés en montant les OSD sur les bons disques (le 1er secteur du disque contient l'id de l'OSD)[44] et recouvré l'ensemble des pools. Depuis nous avons corrigé la règle dans la *crushmap* et écrit un script pour vérifier désormais la bonne répartition des groupes de placement sur les trois salles.

5 Perspectives et évolutions

5.1 Remplacer un NAS FluidFS avec du Ceph et/ou du vSAN ?

Actuellement nous sommes en phase d'études pour le remplacement de notre NAS DELL/FluidFS[45] qui s'appuie sur Compellent.

Le protocole NFS est en cours de remplacement par CephFS (dont le support est natif dans Linux).

En ce qui concerne la partie SMB/CIFS nous envisageons trois pistes.

La première, part du constat que nos deux infrastructures *HCI* offrent des capacités de stockage distribué. vSAN intègre nativement un service de fichier *NFS/SMB* : vSAN File Service[46]. La

42 Taille maximale d'un paquet pouvant être transmis en une seule fois (sans fragmentation) sur une interface réseau

43 Corosync Cluster Engine est un système de communication en groupe

44 Exemple pour retrouver l'identifiant d'un OSD (ici 19) mappé sur le périphérique /dev/dm-2 :

```
dd if=/dev/dm-2 bs=512 count=1 | hexdump -C
[…]
000001e0  6f 61 6d 69 02 00 00 00  31 39 04 79 8c 41 00 00  |oami....19.y.A..|
```

45 Solution de stockage NAS clusterisé qui présente aux clients un serveur de fichiers de type NFS ou SMB/CIFS

46 VSan Filer Service : https://core.vmware.com/resource/vsan-file-services

seconde consiste à faire la même chose sur Proxmox à partir de conteneurs LXC[47] exposant un service Samba clusterisé (*CTDB*), qui accèdent nativement au stockage CephFS sous-jacent. La dernière se veut hybride, à base de *Windows Cluster Failover* [21] accédant à des disques partagés directement attachés aux *VM*. Sous la forme de RBD sous Proxmox et en VMDK partagée sous vSAN.

La solution WSFC fonctionne bien sous vSAN, mais pas sous Proxmox. En effet la validation d'un cluster WSFC nécessite une gestion de la réservation persistante d'un disque *SCSI* (SCSI-3 PR). Même si les RBD sous Ceph sont en mode partagé[48], il n'y a pas de gestion des instructions de persistance au niveau ceph-rbd[49]. Nous avons également tenté d'exposer les disques RBD à travers une gateway iSCSI. Les *LUN* sont bien vues par les *VM*, mais dès qu'on les passe en mode partagé[50], le portail *iSCSI* devient instable. Ces configurations ne sont pas prévues. Il serait techniquement possible d'utiliser DFS-R pour répliquer les volumes au niveau du cluster WSFC. Mais déporter la gestion des réplicas aussi haut dans le cluster n'est pas souhaitable en termes de performance et de résilience.

La solution vSAN File service s'avère peu souple pour nos usages. Il est possible qu'étant donné la jeunesse du service, cela soit amené à s'améliorer. D'autre part, compte tenu des coûts de licence, adosser un service Filer à vSAN peut s'avérer luxueux.

La solution Samba clusterisé sur CephFS, bien que plus complexe semble être ce qui répond le plus à nos besoins en termes de possibilités et de souplesse. CephFS est un système de fichier POSIX compatible avec les contraintes de CTDB (*lock*, *byte range lock* et contention). Il permet la gestion des ACL, des attributs étendus, des quotas et des snapshots de manière native, ce qui rend aisé son intégration dans Samba.

5.2 Réinternalisation et nouveaux services

Depuis la mise en place des *HCI*, le service M*oodle* (sorti de notre périmètre pendant 9 mois) a été réinternalisé : les données sont stockées sur CephFS auxquelles accèdent les frontaux clusterisés sur les deux *HCI*. Nous avons étendu et ouvert à d'autres entités notre offre de stockage et d'hébergement de *VM*. Nous allons internaliser la future plateforme de *Learning Analytics* couplée à la plateforme d'apprentissage *Moodle* (analyse de l'apprentissage) et ses besoins gourmands.

Nous souhaitons également déployer la passerelle objet RADOS Gateway Ceph S3/Swift afin de pouvoir proposer du stockage objet à Nuxeo, l'application de GED[51] opérée par la DSI.

6 Conclusion

Les différents incidents ont amené le taux de disponibilité du cluster Proxmox/Ceph à 99,87 % sur deux ans et celui de vSAN/vSphere à 99,86 % sur un an. Toutefois, les déconvenues sur l'une ou l'autre des infrastructures, liées à des erreurs de configuration ou des bugs désormais résolus ne remettent pas en cause le choix d'exploiter les deux qui sont performantes, stables et fiables. Nous allons contracter du support externe spécifiquement pour Ceph en plus de Proxmox, comme pour vSphere, vSAN et vCenter.

Il nous faut finaliser l'automatisation de la migration des *VM* entre les deux infrastructures, notamment de Proxmox vers vSphere.

47 Linux Container : système de virtualisation, utilisant l'isolation comme méthode de cloisonnement au niveau du système d'exploitation.

48 L'option –image-shared spécifie que l'image RBD peut-être utilisée par plusieurs clients

49 SCSI Persistent Reservations n'est pas encore supporté dans *rbd-wnbd* (RBD windows) : https://docs.ceph.com/en/latest/rbd/rbd-windows

50 La suppression des fonctionnalités *exclusive-lock*, *object-map*, *fast-diff* n'a pas été suffisante : https://docs.ceph.com/en/latest/man/8/rbd

51 Gestion Électronique des Documents

Nous remplacerons nos baies de stockage hétérogènes dédiées aux sauvegardes par un ou plusieurs petits clusters Ceph, nous renforcerons la sauvegarde de notre serveur de fichier CephFS (et ses *snapshots*) avec cephfs-mirror[52].

Nous envisageons d'appliquer les mêmes mesures de filtrage des *VM* sur vSphere que sur Proxmox, après avoir acquis la maîtrise de NSX et également son intégration dans *lsfw* [22], l'outil de la DSI pour tester les règles de pare-feu distribués sur un réseau, comme nous l'avons fait avec les *iptables* gérées par Proxmox.

Comme on a vu pour le NAS, si on veut bénéficier des possibilités natives de stockage distribuée de l'une ou l'autre de nos HCI, les contraintes techniques nous amènent à faire des choix.

Enfin, en apportant plus de solutions, Ceph nous a également demandé plus de travail, mais nous sommes parvenus à prendre en main les différentes technologies, qui répondent à nos besoins actuels et aux nouveaux projets engagés : vSphere est incontournable et Ceph nous apporte l'élasticité qui nous manquait au niveau stockage.

52 Démon permettant la mise en miroir asynchrone des snapshots du système de fichiers CephFS entre des clusters Ceph

Annexe

Fonctionnalités	Proxmox	vSphere
Virtualisation	OUI	
Utilisation de stockage distribué	OUI (avec Ceph)	OUI (avec vSAN)
Infrastructure Hyper-Convergée	OUI (avec Ceph)	OUI (avec vSAN)
Support tous OS pour les VM	OUI (Windows, Linux, *BSD)	
HA pour les VM	OUI (redémarrage VM en cas de crash de l'hôte)	
Snapshots de VM	OUI	
Migration à chaud des VM	OUI	
Migration à chaud du stockage	OUI	
Gestion de templates, de clones	OUI	
Firewall (VM, hôtes et clusters)	OUI (sans licence)	OUI (licence supplémentaire)
DRS	NON	OUI
Console graphique de gestion	OUI (intégrée sur tous les nœuds)	OUI (VM dédiée)
Disponibilité d'un CLI, API	OUI	
Sauvegarde enveloppe VM	OUI	NON
Sauvegarde contenu VM	OUI	NON
Licence	Libre (support payant possible)	Payantes et chères (vSphere, vSAN, vCSA NSX, Tanzu-Kubernetes)
Back-ends de stockage	NFS, SMB/CIFS, GlusterFS, RBD, CephFS, ZFS, iSCSI	NFS, vSAN, iSCSI, FC
Exigence sur le matériel	Aucune	Très fortes (matrice de compatibilité exigeante)
Conteneurs	LXC (sans licence)	Kubernetes (licence supplémentaire)

Figure 6:tableau comparatif entre les fonctionnalités de Proxmox et vSphere

Types de serveurs	DELL R740xd
Processeurs	2 x Intel(R) Xeon(R) Gold 6254 CPU @ 3.10GHz (36 threads)
RAM	768 Go (12 × 64 Go)
Disque OS	Contrôleur BOSS avec 2 cartes M.2 en RAID1
Carte réseau	2 x Intel X710 10 Gb/s SFP+ (4 ports)
Carte RAID	PERC H730p
SSD	19 x 1,92 To SSD SAS Read Intensive

Figure 7 : Spécification matérielle des serveurs hyper-convergés

Types de serveurs	DELL R440
Processeurs	2 × : Intel(R) Xeon(R) Bronze 3204 CPU @ 1.90GHz (6 threads)
RAM	128 Go (8 × 16 Go)
Disque OS	Contrôleur BOSS avec 2 cartes M.2 en RAID1
Carte réseau	Broadcom Gigabit Ethernet BCM5719 (6 ports)

Figure 8 : Spécification matérielle des moniteurs Ceph

Types de serveurs	DELL R440
Processeurs	2 x : Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz (24 threads)
RAM	384 Go (12 × 32Go)
Disque OS	Contrôleur BOSS avec 2 cartes M.2 en RAID1
Carte réseau	Broadcom Adv. Dual 10G SFP+ (2 ports)

Figure 9 : Spécification matérielle du serveur witness vSphere/vSAN

Types de serveurs	DELL R730xd
Processeurs	2 x Intel(R) Xeon(R) CPU E5-2603 v4 @ 1.70GHz (6 threads)
RAM	128 Go (8 × 16 Go)
Disque OS	2 x 200 Go SSD SATA Write Intensive
Carte réseau	2 x Broadcom 57840S 10 Gb/s SFP+ (4 ports)
Carte RAID	PERC H730p
SSD	3 x 200 Go SSD SAS Write Intensive
HDD	13 x 8 To HDD NL-SAS 7,2k

Figure 10 : Spécification matérielle des serveurs du cluster de sauvegarde Proxmox/Ceph

IOPS (Entrées/Sorties par seconde) **Options fio :** -fichier de 4 Go, -blocks de 4 Ko, -profondeur iodepth de 128	**Lectures/Écritures aléatoires (75 % de lectures)**		**Lectures aléatoires IOPS**	**Écritures aléatoires IOPS**
	IOPS en lecture	**IOPS en écriture**		
HCI Proxmox/Ceph	54 000	18 000	94 900	40 100
HCI vSphere/vSAN	32 500	10 800	88 200	18 900
Performances Ceph vs vSAN	+66,15%	+66,67%	+7,6%	+112,17%

Figure 11 : Tests de performances du stockage en lectures/écritures aléatoires avec fio

Bibliographie

[1] Guenael SANCHEZ - Gilles BRUNO - Pascal PRALY. WINTER : virtualisation hyperconvergée Full-Flash multi-site (retour d'expérience sur 1 an). Dans Actes du congrès JRES2017, Nantes, novembre 2017. https://conf-ng.jres.org/2017/document_revision_1923.html?download

[2] David BERCOT, Philippe DUBRULLE, Benjamin GIRARD, Jean-Pierre FEUILLERAT, Olivier LENORMAND et David ROUSSE. Mise en place d'une nouvelle architecture hyper-convergée. Dans Actes du congrès JRES2017, Nantes, novembre 2017. https://conf-ng.jres.org/2017/document_revision_1658.html?download

[3] James S. PLANK : Erasure Codes for Storage Systems, A Brief Primer. Dans Vol.38 n°6 du magazine ;login: de l'association USENIX, pages 44 à 50, décembre 2013. https://www.usenix.org/system/files/login/issues/1312_login.pdf

[4] Yann Dupont. Stockage distribué : retour d'expérience avec CEPH. Dans Actes du congrès JRES2013, Montpellier, décembre 2013. https://2013.jres.org/archives/48/paper48_article.pdf

[5] Stéphane DUGRAVOT, Frédéric NASS. Stockage des données volumineuses de la recherche à l'Université de Lorraine. Dans Actes du congrès JRES2017, Nantes, novembre 2017. https://conf-ng.jres.org/2017/document_revision_2979.html?download

[6] Philippe SABY, Bruno BUISSON. Une infrastructure mutualisée et recyclée avec Ceph. Dans Actes du congrès JRES2017, Nantes, novembre 2017. https://conf-ng.jres.org/2017/document_revision_1952.html?download

[7] Frédéric NASS. Un Zimbra surchargé ? Passez à l'objet. Dans Actes du congrès JRES2017, Nantes, novembre 2017.https://conf-ng.jres.org/2017/document_revision_1632.html?download

[8] Marianne LOMBARD. Cloud privé pour la recherche en informatique : Openstack à l'Inria Saclay. Dans Actes du congrès JRES2017, Nantes, novembre 2017. https://conf-ng.jres.org/2017/document_revision_2975.html?download

[9] Luc DIDRY, Pierre-Yves GOSSET. Quelle infrastructure pour dégoogliser Internet ? Dans Actes du congrès JRES2017, Nantes, novembre 2017. https://conf-ng.jres.org/2017/document_revision_1223.html?download

[10] Yann Dupont. CEPH Retour d'expériences (3 ans, et après ?). Dans Actes du congrès TutoJRES n°18 : le stockage distribué, Lyon CCIN2P3, mai 2016. https://archives.jres.org/_media/tuto/tuto18/tutojres18_ceph-yd.pdf

[11] Pierre BLONDEAU, Davy GIGAN. CEPH. Dans Actes du congrès TutoJRES n°18 : le stockage distribué, Lyon CCIN2P3, mai 2016. https://archives.jres.org/_media/tuto/tuto18/tutojres18_ceph.pdf

[12] https://documentation.suse.com/ses/7/html/ses-all/deploy-additional.html

[13] https://docs.ceph.com/en/latest/rbd/iscsi-overview/#

[14] https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/5/html/block_device_guide/the-ceph-iscsi-gateway

[15] https://docs.ceph.com/en/latest/rbd/iscsi-initiators/

[16] Thomas DOLMAZON. Retour d'expérience sur l'utilisation de Proxmox. Dans Actes du congrès JRES2017, Nantes, novembre 2017. https://conf-ng.jres.org/2017/document_revision_2212.html?download

[17] https://www.cisco.com/c/en/us/support/switches/nexus-93180yc-ex-switch/model.html

[18] https://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsan

[19] Sage A. Weil; Scott A. Brandt; Ethan L. Miller; Carlos Maltzahn. CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data. Dans Actes de la conférence ACM/IEEE Conference on Supercomputing. Tampa, FL, USA, novembre 2006. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4090205

[20] https://groupes.renater.fr/sympa/arc/ceph/2022-01/msg00008.html

[21] https://docs.microsoft.com/en-us/windows-server/failover-clustering/failover-clustering-overview

[22] Patrick LAMAIZIERE. Lsfw, outil de test de règles de pare-feu distribués sur un réseau. Dans Actes du congrès JRES2011, Toulouse, novembre 2011. https://conf-ng.jres.org/2011/document_revision_1341.html?download