PANNE ELECTRIQUE

Bonjour,
Ce matin à 7h23, nous avons eu une panne majeure sur notre site de Strasbourg (SBG) : une coupure électrique qui a mis dans le noir nos 3 datacentres SBG1, SBG2 et SBG4 durant 3h30. Le pire scénario qui puisse nous arriver.

Le site de SBG est alimenté par une ligne électrique de 20KVA composée de 2 câbles qui délivrent chacun 10MVA. Les 2 câbles fonctionnent ensemble, et sont connectés à la même source et sur le même disjoncteur chez ELD (Strasbourg Électricité Réseaux). Ce matin, l'un des 2 câbles a été endommagé et le disjoncteur a coupé l'alimentation des datacentres.

Le site SBG est prévu pour fonctionner, sans limite de temps, sur les groupes électrogènes. Pour SBG1 et SBG4, nous avons mis en place, un premier système de 2 groupes électrogènes de 2MVA chacun, configurés en N+1 et en 20KV. Pour SBG2, nous avons mis en place 3 groupes en N+1 de 1.4MVA chacun. En cas de coupure de la source externe, les cellules haute tension sont reconfigurées automatiquement par un système de bascule motorisé. En moins de 30 secondes, les datacentres SBG1, SBG2 et SBG4 sont ré-alimentés en 20KV. Pour permettre toutes ces bascules sans couper l'alimentation électrique des serveurs, nous disposons d'onduleurs (UPS) sachant fonctionner sans aucune alimentation durant 8 minutes.

Ce matin, le système de basculement motorisé n'a pas fonctionné. L'ordre de démarrage des groupes n'a pas été donné par l'automate. Il s'agit d'un automate NSM (Normal Secours Motorisé), fournit par l'équipementier des cellules haute-tension 20KV. Nous sommes en contact avec lui, afin de comprendre l'origine de ce dysfonctionnement. C'est toutefois un défaut qui aurait dû être détecté lors des tests périodiques de simulation de défaut sur la source externe. Le dernier test de reprise de SBG sur les groupes date de la fin du mois mai 2017. Durant ce dernier test, nous avons alimenté SBG uniquement à partir des groupes électrogènes durant 8H sans aucun souci et chaque mois nous testons les groupes à vide. Et malgré tout, l'ensemble de ce dispositif n'a pas suffi aujourd'hui pour éviter cette panne.

Vers 10h, nous avons réussi à basculer les cellules manuellement et nous avons recommencé à alimenter le datacentre à partir des groupes électrogènes. Nous avons demandé à ELD de bien vouloir déconnecter le câble défectueux des cellules haute tension et remettre le disjoncteur en marche avec 1 seul des 2 câbles, et donc limité à 10MVA. La manipulation a été effectuée par ELD et le site a été ré-alimenté vers 10h30. Les routeurs de SBG ont été joignables à partir de 10h58.

Depuis, nous travaillons, sur la remise en route des services. Alimenter le site en énergie permet de faire redémarrer les serveurs, mais il reste à remettre en marche les services qui tournent sur les serveurs. C'est pourquoi chaque service revient progressivement depuis 10h58. Notre système de monitoring nous permet de connaitre la liste de serveurs qui ont démarré avec succès et ceux qui ont encore un problème. Nous intervenons sur chacun de ces serveurs pour identifier et résoudre le problème qui l'empêche de redémarrer.

A 7h50, nous avons mis en place une cellule de crise à RBX, où nous avons centralisé les informations et les actions de l'ensemble des équipes. Un camion en partance de RBX a été chargé de pièces de rechange pour SBG. Il est arrivé à destination vers 17h30. Nos équipes locales ont été renforcées par des équipes du datacentre de LIM en Allemagne et de RBX, ils sont tous mobilisés sur place depuis 16H00. Actuellement, plus de 50 techniciens travaillent à SBG pour remettre tous les services en route. Nous préparons les travaux de cette nuit et, si cela était nécessaire, de demain matin.

Prenons du recul. Pour éviter un scénario catastrophe de ce type, durant ces 18 dernières années, OVH a développé des architectures électriques capables de résister à toutes sortes d'incidents électriques. Chaque test, chaque petit défaut, chaque nouvelle idée a enrichi notre expérience, ce qui nous permet de bâtir aujourd'hui des datacentres fiables.

Alors pourquoi cette panne ? Pourquoi SBG n'a pas résisté à une simple coupure électrique d'ELD ? Pourquoi toute l'intelligence que nous avons développée chez OVH, n'a pas permis d'éviter cette panne ?

La réponse rapide : le réseau électrique de SBG a hérité des imperfections de design liées à la faible ambition initialement prévue pour le site.

La réponse longue :
En 2011, nous avons planifié le déploiement de nouveaux datacentres en Europe. Pour tester l'appétence de chaque marché, avec de nouvelles villes et de nouveaux pays, nous avons imaginé une nouvelle technologie de déploiement de datacentres, basée sur les containers maritimes. Grâce à cette technologie, développée en interne, nous avons voulu avoir la souplesse de déployer un datacentre sans les contraintes de temps liées aux permis de construire. A l'origine, nous voulions avoir la possibilité de valider nos hypothèses avant d'investir durablement dans un site.

C'est comme ça que début 2012, nous avons lancé SBG avec un datacentre en containers maritimes : SBG1. Nous avons déployé 8 containers maritimes et SBG1 a été opérationnel en seulement 2 mois. Grâce à ce déploiement ultra rapide, en moins de 6 mois nous avons pu valider que SBG est effectivement un site stratégique pour OVH. Fin 2012,

nous avons décidé de construire SBG2 et en 2016, nous avons lancé la construction de SBG3. Ces 2 constructions n'ont pas été faites en containers, mais ont été basées sur notre technologie de « Tour » : la construction de SBG2 a pris 9 mois et SBG3 sera mis en production dans 1 mois. Pour pallier aux problèmes de place début 2013, nous avons construit très rapidement SBG4, l'extension basée encore sur les fameux containers maritimes.

Le problème est qu'en déployant SBG1 avec la technologie basée sur les containers maritimes, nous n'avons pas préparé le site au large scale. Nous avons fait 2 erreurs :
1) nous n'avons pas remis le site SBG aux normes internes qui prévoient 2 arrivées électriques indépendantes de 20KV, comme tous nos sites de DCs qui possèdent plusieurs doubles arrivées électriques. Il s'agit d'un investissement important d'environ 2 à 3 millions d'euros par arrivée électrique, mais nous estimons que cela fait partie de notre norme interne.
2) nous avons construit le réseau électrique de SBG2 en le posant sur le réseau électrique de SBG1, au lieu de les rendre indépendant l'un de l'autre, comme dans tous nos datacentres. Chez OVH, chaque numéro de datacentre veut dire que le réseau électrique est indépendant d'un autre datacentre. Partout sauf sur le site de SBG.

La technologie basée sur les containers maritimes n'a été utilisée que pour construire SBG1 et SBG4. En effet, nous avons réalisé que le datacentre en containers n'est pas adapté aux exigences de notre métier. Avec la vitesse de croissance de SBG, la taille minimale d'un site est forcément de plusieurs datacentres, et donc d'une capacité totale de 200.000 serveurs. C'est pourquoi, aujourd'hui, pour déployer un nouveau datacenter, nous n'utilisons plus que 2 types de conceptions largement éprouvées et prévues pour le large scale avec de la fiabilité :
1) la construction de tours de 5 à 6 étages (RBX4, SBG2-3, BHS1-2), pour 40.000 serveurs.
2) l'achat des bâtiments (RBX1-3,5-7, P19, GRA1-2, LIM1, ERI1, WAW1, BHS3-7, VIH1, HIL1) pour 40.000 ou 80.000 serveurs.

Même si l'incident de ce matin a été causé par un automate tiers, nous ne pouvons nous dédouaner de la responsabilité de la panne. A cause du déploiement initial basé sur les containers maritimes, nous avons un historique à rattraper sur SBG pour atteindre le même niveau de normes que sur les autres sites d'OVH.

Cet après-midi, nous avons décidé du plan d'actions suivant :
1) la mise en place de la 2ème arrivée électrique, totalement séparée, de 20MVA ;
2) la séparation du réseau électrique de SBG2 vis-à-vis de SBG1/SBG4, ainsi que la séparation du futur SBG3 vis-à-vis de SBG2 et SBG1/SBG4;
3) la migration des clients de SBG1/SBG4 vers SBG3 ;
4) la fermeture de SBG1/SBG4 et la désinstallation des containers maritimes.

Il s'agit d'un plan d'investissement de 4-5 millions d'euros, que nous mettons en route dès demain, et qui, nous l'espérons, nous permettra de restaurer la confiance de nos clients envers SBG et plus largement OVH.

Les équipes sont toujours en train de travailler sur la remise en route des derniers clients impactés. Une fois l'incident clos, nous appliquerons les SLA prévus dans nos contrats.

Nous sommes profondément désolés pour la panne générée et nous vous remercions des encouragements que vous nous témoignez durant cet incident.

Amicalement
Octave

PANNE SUR LE RESEAU OPTIQUE

Bonjour,
Ce matin, nous avons eu un incident sur le réseau optique qui interconnecte notre site de Roubaix (RBX) avec 6 des 33 points de présence (POP) de notre réseau : Paris (TH2 et GSW), Francfort (FRA), Amsterdam (AMS), London (LDN), Bruxelles (BRU).

Le site RBX est connecté à travers 6 fibres optiques à ces 6 POP : 2x RBX<>BRU, 2x RBX<>LDN, 2x RBX<>Paris (1x RBX<>TH2 et 1x RBX<>GSW). Ces 6 fibres optiques sont connectées aux systèmes de nœuds optiques qui permettent d'avoir 80 longueurs d'onde de 100Gbps sur chaque fibre optique.

Pour chaque 100G connectés aux routeurs, nous utilisons 2 chemins optiques qui sont géographiquement distincts. En cas de coupure de fibre optique, le fameux « coup de pelleteuse », le système se reconfigure en 50ms et tous les liens restent UP. Pour connecter RBX aux POP, nous avons 4.4Tbps de capacité, 44x100G : 12x 100G vers Paris, 8x100G vers London, 2x100G vers Bruxelles, 8x100G vers Amsterdam, 10x100G vers Frankfurt, 2x100G vers DC GRA et 2x100G vers DC SBG.

A 8h01, d'un coup, l'ensemble des liens 100G, les 44x 100G, ont été perdus. Étant donné le système de redondance que nous avons mis en place, l'origine du problème ne pouvait pas être la coupure physique de 6 fibres optiques

simultanément. Nous n'avons pas pu faire les diagnostiques sur les châssis à distance car les interfaces de management étaient figées. Nous avons été obligés d'intervenir directement dans les salles de routage, pour faire les manipulations sur les châssis : déconnecter les câbles entre les châssis puis faire redémarrer le système et enfin seulement faire les diagnostiques avec l'équipementier. Les tentatives de redémarrage du système ont pris beaucoup de temps, car chaque châssis a besoin de 10 à 12 minutes pour démarrer. C'est la principale raison de la durée de l'incident.

Le diagnostique : Toutes les cartes transpondeurs que nous utilisons, ncs2k-400g-lk9, ncs2k-200g-cklc, sont passées en état « standby ». L'une des origines possible d'un tel état est la perte de configuration. Nous avons donc récupéré le backup et remis en place la configuration, ce qui a permis au système de reconfigurer toutes les cartes transpondeurs. Les 100G dans les routeurs sont revenus naturellement et la connexion de RBX vers les 6 POP a été rétablie à 10h34.

Il s'agit clairement d'un bug software sur les équipements optiques. La base de données avec la configuration est enregistrée 3 fois et copiée sur 2 cartes de supervision. Malgré toutes ces sécurités, la base a disparu. Nous allons travailler avec l'équipementier pour trouver l'origine du problème et les aider à fixer le bug. Nous ne remettons pas en cause la confiance avec l'équipementier, même si ce type de bug est particulièrement critique. L'uptime est une question de design qui prend en compte tous les cas de figure, y compris quand plus rien ne marche. Le mode parano chez Ovh doit être poussé encore plus loin dans l'ensemble de nos designs.

Les bugs ça peut exister, les incidents qui impactent nos clients non. Il y a forcement une erreur chez Ovh puisque malgré tous les investissements dans le réseau, dans les fibres, dans les technologies, nous venons d'avoir 2 heures de downtime sur l'ensemble de nos infrastructures à Roubaix.

L'une des solutions est de créer 2 systèmes de nœuds optiques au lieu d'un seul. 2 systèmes, cela veut dire 2 bases de données et donc en cas de perte de la configuration, un seul système est en panne. Si 50% des liens passent par l'un des systèmes, aujourd'hui, nous aurions perdu 50% de la capacité mais pas 100% de liens. C'est l'un des projets que nous avons commencé il y a 1 mois, les châssis ont été commandés et nous allons les recevoir dans les prochains jours. Nous pourrons commencer les travaux de configuration et migration sous 2 semaines. Vu l'incident d'aujourd'hui, ce projet devient prioritaire, pour l'ensemble de nos infrastructures, tous les DCs, tous les POPs.

Dans le métier de fournisseur des infrastructures Cloud, seul ceux qui sont paranos durent. La qualité de service est une conséquence de 2 éléments. Tous les incidents anticipés « by design ». Et les incidents où nous avons appris de nos erreurs. Cet incident là nous amène à mettre la barre encore plus haut pour s'approcher du risque zéro.

Nous sommes sincèrement désolés pour les 2H33 minutes de downtime sur le site RBX. Dans les prochains jours, les clients impactés vont recevoir un email pour déclencher l'application des engagements SLA.

Amicalement
Octave