Equilibrage des charges : l'incontournable d'une infrastructure haute disponibilité

Le marché de l'équilibrage des charges (load balancing) devrait atteindre 5 milliards de dollars d'ici 2023. Cette croissance est portée par des tendances telles que le haut débit mobile, les environnements multicloud et hybrides, la virtualisation, le télétravail et l'utilisation des appareils personnels.

Par conséquent, la haute disponibilité est devenue incontournable pour des applications aussi essentielles que le PGI (Progiciel de Gestion Intégré), les communications et les systèmes collaboratifs, ou encore les infrastructures VDI (Virtual Desktop Infrastructure).

La haute disponibilité, un enjeu indispensable

La haute disponibilité désigne la capacité d'un système, ou de l'un de ses composants, à rester constamment opérationnel pendant de longues périodes. Pour cela, les services informatiques appliquent certaines règles comme la répartition de la charge, la redondance ou la tolérance aux pannes *(FailOver)* pour assurer un fonctionnement continu et une correction rapide de la panne.

Cela s'applique à tous les éléments du datacenter : pour la haute disponibilité, pour les applications, pour l'équilibreur des charges ou le contrôleur de mise à disposition d'applications (ADC) qui gère le trafic réseau, aussi bien dans un centre de données que dans un environnement qui en regroupe plusieurs.

La mise en place d'un environnement haute disponibilité commence par l'identification et l'élimination de tous les maillons faibles de l'infrastructure risquant de déclencher une interruption de service. Par exemple, en déployant des composants redondants pour assurer la tolérance de panne en cas de défaillance de l'un des appareils. L'équilibrage des charges, qu'il soit réalisé au moyen d'un appareil unique ou d'un contrôleur ADC, facilite ce processus en suivant l'état de santé des serveurs, en détectant les pannes éventuelles et en redirigeant le trafic de façon à assurer un service ininterrompu.

Si la tolérance de panne est essentielle pour les serveurs, l'architecture haute disponibilité doit également prendre en compte la couche d'équilibrage des charges. S'il devient impossible d'exécuter cette fonction de façon efficace, les serveurs risquent la saturation, ce qui a non seulement un impact sur leur bon fonctionnement mais affecte également les performances et la disponibilité des applications. Par conséquent, la redondance est aussi importante pour l'équilibrage des charges ou le contrôleur ADC que pour tout autre composant du datacenter.

Garder une certaine tolérance aux pannes

Comme pour un cluster de serveurs haute disponibilité, les équilibreurs de charges et les contrôleurs ADC peuvent être déployés de différentes façons pour assurer la haute disponibilité :

1. **Actif-passif** – Il s'agit de la configuration la plus courante. Le modèle actif-veille inclut une instance redondante de chaque contrôleur ADC qui prend le relais dès que le nœud principal fait défaut. Chaque contrôleur ADC actif peut être configuré différemment, mais chaque paire actif-passif doit utiliser les mêmes paramètres.
2. **Actif-actif** – Avec ce modèle, plusieurs contrôleurs ADC configurés de la même façon sont déployés pour l'utilisation ordinaire. En cas de panne de l'un des nœuds, son trafic est repris par au moins l'un des nœuds restants et les charges sont équilibrées pour assurer la continuité du service. Cette approche suppose que le cluster dispose d'une capacité suffisante pour assurer un fonctionnement normal en cas de défaillance de l'un des contrôleurs ADC.
3. **N+1** – Cette solution permet d'assurer la redondance à moindre coût par rapport au modèle actif-passif, car dans ce cas un contrôleur ADC n'est mis en ligne qu'en cas de défaillance du contrôleur ADC principal.

Dans chaque cas, un basculement rapide permet d'assurer la tolérance aux pannes et la reprise après sinistre pour la fonction d'équilibrage des charges, de sorte que les performances et la disponibilité de l'application ne soient pas compromises par la panne. La tolérance de panne et la gestion du trafic sont généralement traitées par l'une des versions du protocole de redondance de routeur virtuel (VRRP).

Principales caractéristiques de haute disponibilité pour l'équilibrage des charges ou le contrôleur ADC

Il est à la fois nécessaire de garantir la haute disponibilité d'un contrôleur ADC, mais également d'assurer que ce dernier offre bien la haute disponibilité au trafic des applications gérées. En cas de panne d'un serveur, le contrôleur ADC peut réacheminer le trafic vers un autre serveur du cluster.

1. **Méthodes d'équilibrage des charges** – Plusieurs méthodes de sélection du serveur sont possibles : *round robin, least connections, weighted round robin, weighted least connections, fastest response*, etc. Le contrôleur ADC doit réunir toutes ces options pour établir la configuration la mieux adaptée à l'environnement et aux priorités établies.
2. **Suivi de l'état de santé** – Afin d'assurer un basculement rapide avec aussi peu de temps d'indisponibilité que possible, l'état de santé du serveur doit être constamment évalué en fonction de différents indicateurs, notamment :

1. Le nombre total d'octets entrant sur le serveur et en sortant pendant une période donnée
2. Le débit du trafic (en Mbit/s) entrant sur le serveur et en sortant pendant une période donnée
3. Le pourcentage du trafic d'erreur par rapport à la plage considérée

4. Le nombre de bonnes connexions SSL

5. La latence moyenne du serveur d'applications par service

6. Le SRTT de latence côté client, max, min et moyenne pendant une période donnée

7. Les vérifications personnalisées de l'état de santé, telles que la mesure du temps de réponse pour des requêtes de base de données spécifiques

Toute période d'indisponibilité peut avoir des conséquences graves pour les entreprises. Les pertes liées à ces périodes sont estimées à 780 000 £ par semaine (soit 873 000 €) pour une entreprise d'environ 10 000 personnes. Les pertes directes sont substantielles et justifient à elles seules une solution haute disponibilité.

Afin, d'assurer la disponibilité complète des applications et ou de services de production, une fois la tolérance de panne et la répartition de la charge prise en compte, il convient également de considérer les aspects de Cybersécurité. En effet, une attaque réseau ou applicative, si elle n'est pas identifiée par la solution ADC, traitera le traffic et répartira la charge de l'attaque, ce qui rendrait à nouveau les applications indisponibles.

Ainsi, il convient de considérer systématiquement des ADC "Next Generation"qui intègrent nativement des fonctions de sécurité Firewall, de Web Application Firewall, de Firewall DNS, de protection AntiDDoS, de Proxification, d'authentification avancée, de déchiffrement/Re chiffrement centralisé des flux TLS/SSL ainsi que d'une administration et d'un reporting centralisé et consolidé tant au niveau du réseau, que des évènements de sécurité divers et de l'état des services applicatifs.

Hormis le coût direct de l'indisponibilité, la continuité métier est également en jeu, aussi bien en termes de perte de réputation que de données, et nécessite également l'implémentation de la haute disponibilité. La réputation est ainsi améliorée, car l'entreprise et sa marque deviennent synonymes de fiabilité par rapport aux concurrents. D'autre part, la réduction du risque de perte de données est essentielle, ne serait-ce que pour éviter les pénalités sévères prévues par le Règlement général sur la protection des données personnelles. Enfin, une infrastructure haute disponibilité permet d'atténuer l'impact négatif des pannes sur les revenus et la productivité.