UNIVERSITÉ DE STRASBOURG

Ceph à la DNum de l'Université de Strasbourg

Pôle ICS <dnum-ics@unistra.fr>

Plan de la présentation

- Contexte
- Infrastructure Ceph
- Fonctionnalités utilisées
- Évolutions envisagées
- Retour d'expérience
- Conclusion

Contexte

- Avant 2015 :
  - Stockage local et baie de stockage traditionel
- 2015 : Projet Industrialisation de la Virtualisation
  - Besoins :
    - Migration à chaud des machines virtuelles
    - Rationalisation de l'espace de stockage
- 2015 : Projet Seafile
  - Déploiement initial sur du NFS
  - Volumétrie qui "scale" facilement
- 2018 : Projet Cloud Unistra
  - Mise en place d'un cloud privé OpenStack
  - Offre de service IaaS et stockage S3

Infrastructure Ceph

- Cluster de production
  - 3 Monitors
  - 33 serveurs d'OSDs
    - 9 Serveurs NVMe : 10*6.4To NVMe
    - 6 Serveurs SSDs : 15*1.8To SSD
    - 12 Serveurs HDD : 18*16To 7.2k SATA + 2*6.4 To NVMe
    - 6 Serveurs HDD : 12*8To 7.2K SATA + 2*1.8To SSD
  - 4 machines virtuelles pour S3/RadosGW
  - 2 machines virtuelles pour le LB de S3/RadosGW

- Cluster de backup
  - 1 Monitor
  - 6 serveurs d'OSDs : 16*8To 7.2k SATA

Infrastructure Ceph

- Version : Red Hat Ceph Storage 4
  - Initialement déployé en version communautaire
  - Migration vers Red Hat car besoin de support
- OS des noeuds : RHEL 8
- Système de déploiement : ceph-ansible
- Évolution futur / passage à Red Hat Ceph Storage 5 :
  - Conteneurisation des daemons
  - Passage à cephadm

Fonctionnalités utilisées

- Rados Block Device (RBD)
  - Stockage bloc des VMs sur OpenStack
  - Deux pools différents en fonction des besoins :
    - Tier 1 / performance:
      - Pool composé de NVMe et de SSDs
      - Réplication x3 : ~ 160 To / 210 To utile
    - Tier 2 / volumétrie :
      - Pool composé de disque mécanique
      - Réplication x3 : ~ 60 To / 160 To utile

Fonctionnalités utilisées

- Rados Gateway (RGW)
  - Utilisation de S3 et Swift :
    - S3 : Offre de service stockage objet
    - Swift : Intégration avec OpenStack
  - Pool unique pour l'ensemble des objets
    - Disques mécaniques
    - Erasure coding : 4+2
    - Failure domain : rack
    - ~ 267 To / 1.6 Po utile
  - Load-balancing : keepalived + HAProxy

Fonctionnalités utilisées

- CephFS
  - Pas d'utilisation de CephFS pour le moment à la DNUM
  - Besoin de mettre à disposition du stockage filesystem via OpenStack Manila :
    - Solution 1 : CephFS natif dans les VMs
      - Sécurité dans le cadre d'une offre multi-tenant ?
    - Solution 2 : Paserelle Ganesha
      - Stabilité et performance à discuter ?
    - Solution 3 : Virtio-FS
      - Solution plébiscité mais pas d'implémentation dans OpenStack Manila

Évolutions envisagées

- RBD Persistent Write Log Cache
  - Cache local persistant directement sur les hyperviseurs
  - Améliorer la latence des IOs pour les applications très sensibles (BDD, etc)

- RadosGW :
  - load-balancing pour l'API S3/Swift
    - Le load-balancer encaisse une forte charge réseau (traffic) et CPU (terminaison SSL)
    - Solution : load-balancing via BGP / ECMP
  - Réplication de la zone sur un site distant
    - En attente de l'implémentation du Dynamic Bucket Index resharding

Retour d'expérience

- Administration parfois un peu complexe qui tend à se simplifier
- Système de stockage qui scale très bien
  - Performance et volumétrie
- Incidents depuis sa mise en place :
  - Lenteurs lié aux nombres de snapshots RBD
  - Perte de données suite à un bug sur les RadosGW
- Pas d'interruptions depuis sa mise en service

Conclusion

- Système de stockage principal pour la DNum de l'Université de Strasbourg
  - Offre de service IaaS et S3
- Volonté de s'éloigner le plus possible des baies traditionnelles
  - Difficulté à trouver un équivalent aujourd'hui pour CIFS à notre échelle
  - Remplacement de NFS par OpenStack Manila
- Satisfait de la solution depuis sa mise en service en 2015

Questions