Retour d'expérience BeeGFS dans GRICAD

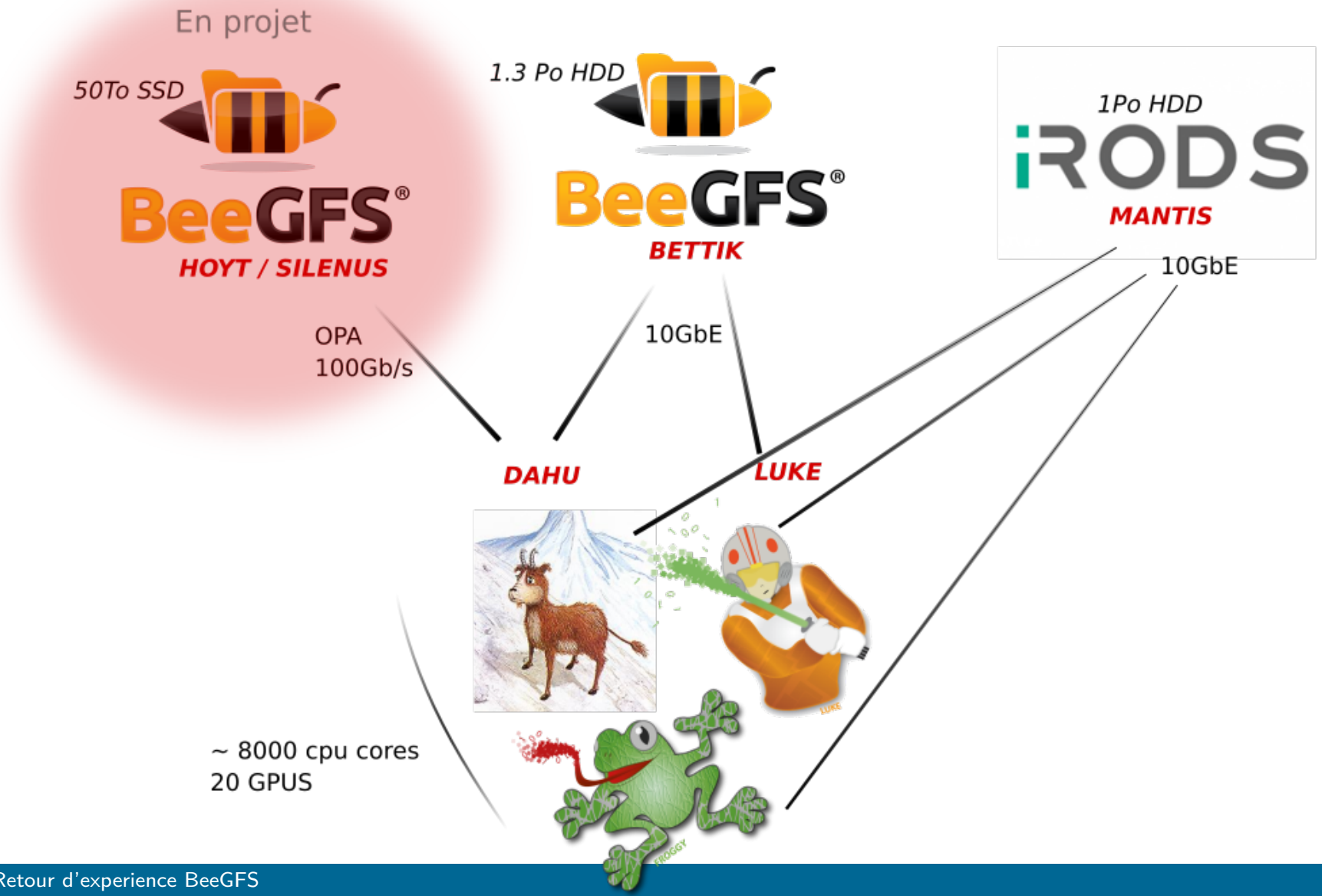
sur l'administration et l'utilisation

Jean-Noel Bouvier (ISTERRE)
Bruno Bzeznik (GRICAD)
Albanne Lecointre (ISTERRE)

4 décembre 2020

GRICAD
GRENOBLE ALPES RECHERCHE
INFRASTRUCTURE DE CALCUL INTENSIF ET DE DONNÉES

Les plateformes de calcul à Gricad



Le projet "Bettik"

En 2017, GRICAD, avec l'aide d'un consortium d'ingénieurs de laboratoires Grenoblois, a connecté ses plate-formes à un système de fichiers distribué performant et évolutif destiné aux données de calcul. D'une capacité de 160To au début du projet, ce système géré sous BeeGFS, fournit aujourd'hui 1.3Po net de stockage de type "workdir", répartis sur 20 serveurs, connectés en 10 GbE depuis plus de 150 noeuds de calcul utilisés par environ 400 utilisateurs actifs.

Le projet "Bettik"

Le 20/11/2020, dans /bettik:

- **Meta-data: 4 serveurs, 32 SSD**
- **Data: 16 serveurs, 224 HDD**
- **Interco 10GbE**
- **1.1 Po used / 1.3 Po total**
- **300 millions de fichiers**
- **301 utilisateurs ont plus de 100 fichiers**
- **8 en ont plus de 10 millions**
- **On voit souvent des I/O avoisinant 10Go/s agrégé**
- **Certains jobs tirent 400000 IOPS**

Le prototype "Hoyt"

Pendant le confinement, en 2020, pour des besoins très spécifiques impliquant des recherches autour de la Covid19, GRICAD a mis en place en urgence un BeeGFS expérimental exploitant les SSD des noeuds de calcul "bac-à-sable" du supercalculateur Dahu afin de franchir des barrières en termes d'IOPS.

Le projet "Silenus"

Fort du succès du prototype **Hoyt**, encore exploité aujourd'hui pour des jobs à fort besoins d'IOPS sur des petits fichiers, GRICAD a décidé de mettre en place une solution complémentaire à **Bettik**, ayant les caractéristiques suivantes:

- Interco sur le réseau Omnipath 100Gb/s de Dahu
- Noeud de meta-data à base de NVMe
- Noeud de data à base de SSD
- Gestion de type "scratch" (effacement des données à la fin des jobs)
- Plateforme initiale de 50 To utiles (1 serveur de meta x4 NVMe et 4 serveurs de Data x 8 SSD 1.6To write intensive)

Monitoring via Elasticsearch/Logstash/Grafana

Voir présentation dédiée "Profiling, suivi et statistiques des jobs de calcul et des accès aux stockages distribués avec ElasticSearch et Grafana"

Bilan de 4 années d'exploitation de BeeGFS

- Une grande satisfaction des admins et des utilisateurs
  - les performances sont au rendez-vous
  - installation facile, grande évolutivité
  - stable
- Le monitoring via un ELK ou équivalent nous parait indispensable pour un système sous forte charge (prévoir une infra conséquente si l'on souhaite conserver les données)
- Le support commercial:
  - bien que tentant dans les périodes difficiles ("mais pourquoi ça rame???") n'a jamais été vraiment nécessaire
  - obligatoire pour certaines fonctionnalités, comme les quotas: nous prévoyons des quotas "informatifs"
- Ne pas négliger la charge de travail : Exemple, pour un Workdir d'1.2 Po exploités par 300 utilisateurs actifs / 4000 coeurs
  $\Longrightarrow$ 1 ETP au début du projet, travail en équipe conseillé, puis 0.2 ETP en régime de croisière
- L'implication de **power users** dans l'équipe d'admins est très bénéfique
- L'homogénéité du hardware et des versions de firmware, même non requise, est un plus non négligeable

Un sacré travail d'équipe!

Merci à...

Jean-Noel Bouvier (Isterre)

Nicolas Gibelin (Gricad)

Anthony Defize (Gricad)

Albanne Lecointre (Isterre)

Patrick Begou (Legi)

Françoise Roch (OSUG)

Et aux nombreux "power users" pour leur patience et collaboration:

Geoffroy Lesur, Arnaud Pladys, Romain Brossier, Ludovic Metivier,...