Une transition vers le système de fichiers Rozo

au Centre de Calcul de l'Université de Strasbourg

Romaric DAVID, Michel RINGENBACH
david@unistra.fr, mir@unistra.fr
Direction Informatique
Tuto Jres - 4 Mai 2016

UNIVERSITÉ DE STRASBOURG

This talk focuses on an important technological change in the HPC Center of the University of Strasbourg (Unistra)

- Unistra is one of the major universities in France:
  - 48 000 students
  - 4800 employees
  - 3 Nobel prizes since 1987
  - major research institute in many scientific domains
  - ...some of them need HPC
- HPC Center (http://hpc.unistra.fr) serves the whole Alsace Region

UNIVERSITÉ DE STRASBOURG

HPC in Strasbourg: Flops, Bytes....

- Around 350 servers, 5500 cores
- 450 TB of GPFS Storage (on leave)
- 516 TB of RozoFS storage (being populated)
- Many TB of BeeGFS
- 60 GPUs, from Tesla
  M2050 to K80
- 223 Tflops
- More than 250 active users
- More than 150
  softwaremodules

Plan

- HPC in Strasbourg University
- « My simulation is slow »
- Cascade effect
- Conclusion

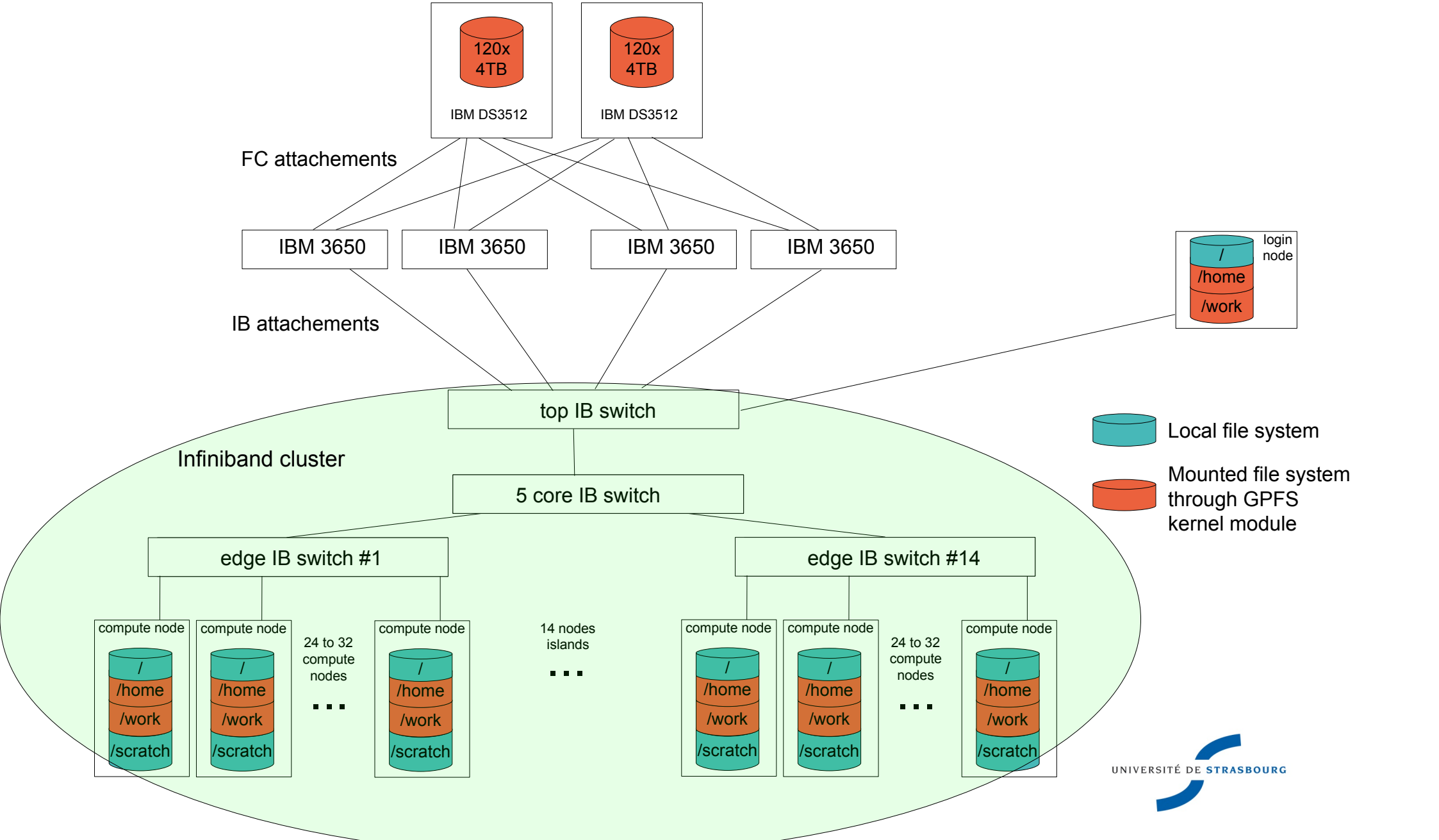
« My simulation is slow »

- Once upon a time...
  - We were challenged by users of the Relion application http://www2.mrc-lmb.cam.ac.uk/relion/index.php/Main_Page, used for reconstruction of 2D or 3D classes in cryo-electron microscopy
  - The execution times on the computing centers were much more longer than in the lab computers
  - Strategically not acceptable for the HPC Center
- At first we went wrong:
  - Compared speed-up against standard experiments of the author of the code (Sjors Scheres)
  - Profiled the application on users data-sets: I/O problems
  - Several hundreds of small files in real Datasets

« My simulation is slow »

- GPFS is the suspect: 1GB/s or more but totally flooded if small files in directories (a surprising big-data effect, made of small datas !)
- GPFS bought in early 2012, first Relion alert : mid 2014
- Other big challenging applications arrived in the meantime

GPFS/Infiniband setup



UNIVERSITÉ DE STRASBOURG

GPFS setup

- Can we scale-up the configuraton ?
- Our GPFS setup in not easy to extend
  - A lot of Fiber-Channel links (redundant direct-attach)
  - No SAN
  - We would have to partition disk space if we were to add more GPFS servers

« My simulation is slow »

- BeeGFS was used to host very-hot data (see previous talks in this TutoJres)
- We are now willing to integrate on-demand BeeGFS for our HPC jobs
- Centralized storage is still an issue : how to provide cost-effective « middle-speed » access to **/home** data ?
- We decided to rely on Ethernet for this... see below!

Plan

- HPC in Strasbourg University
- « My simulation is slow »
- Cascade effect
- Conclusion

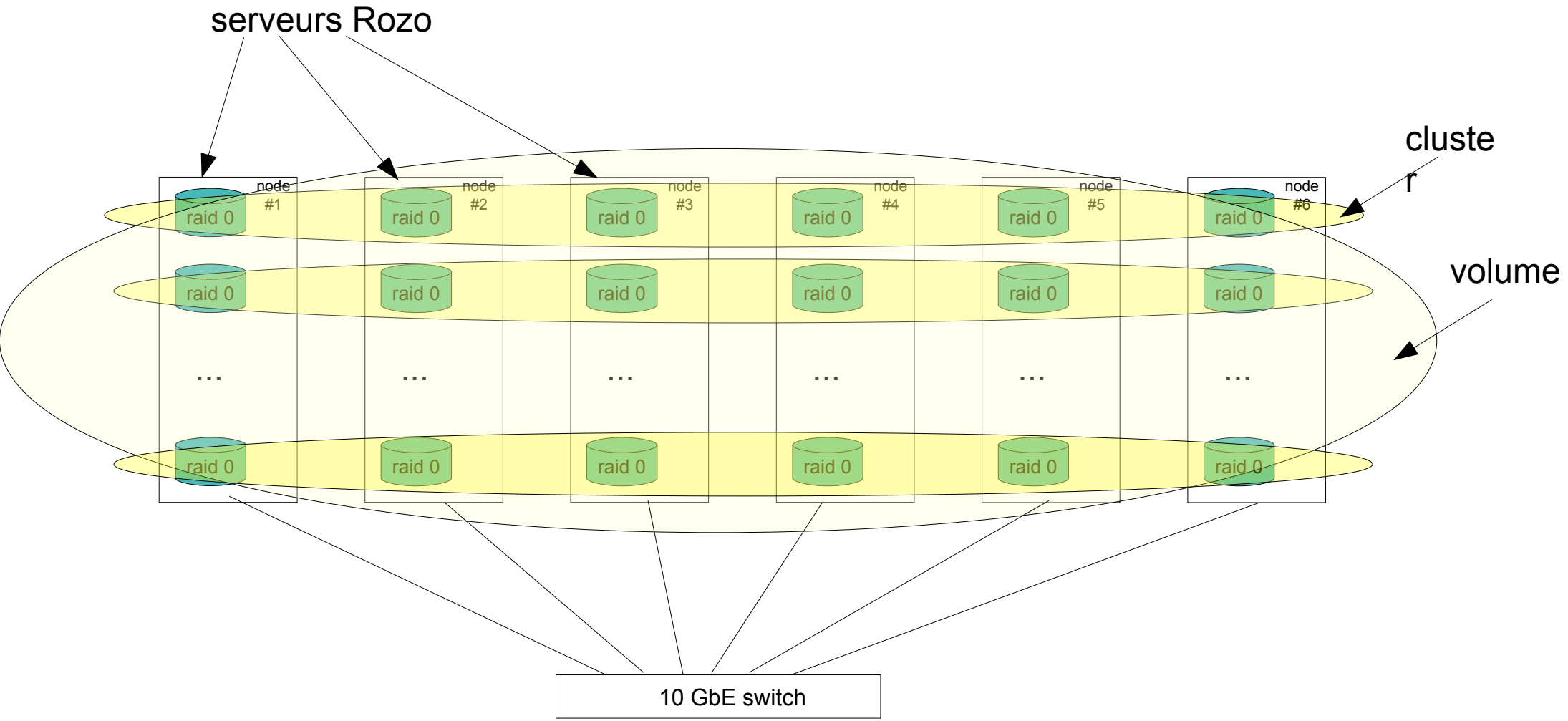
Cascade effect

- **Statement: we want to build upon capacitive drives (7200 RPM, >= 4 TB) + GbE network**
- We tried on-site several filesystems (Thanks to Dell and Rozo Systems)
- We had a long onsite test of RozoFS
  - RozoFS : SDS, based on standard hardware, NFS and *native* mode via fuse
- We compared it with GPFS and BeeGFS using identical benchmarks

RozoFS

- RozoFS : SDS
- Config de test : 4 "projections" par fichier (code à effacement par transformées de Mojette), réparties sur 6 serveurs
  - Coût disque = brut = 1,5x utile
  - Perte possible de 2 serveurs pour 1 fichier
    - Reconstruction sur un autre serveur en cas de perte
- Pas de RAID. Gros RAID très risqués sur gros disques, car les défaillances se produisent souvent en série
- En théorie, N-4 perte de serveurs possible
  - si place suffisante sur 4 serveurs
  - si les reconstructions ont le temps de se faire

RozoFS setup



Procédure de test

- Lors des tests, on accède au filesystem Rozo depuis un grand nombre de noeuds :
  - 128 noeuds sont facilement mobilisables
  - Jobs en cours suspendus
  - noeuds équipés de 1 GbE, mais saturation des liens vers les POC au bout d'un nombre suffisant de noeuds. On arrive à remplir le réseau.
- E/S brutes en volume : dd
- E/S aléatoires : fio (couteau suisse)
- Différentes mesures ("ls -lR" = point de souffrance des utilisateurs)
- Benchs figés au bout d'un moment, pour faciliter la comparaison

Procédure de test

Exemple : fio exécuté sur un cluster de 128 noeuds

```
DEST=rozofs # dellfs|workdir|scratch-beegfs|rozofs : /$DEST/fio/ doit exister
 ND=101
 SCRIPT=/$DEST/fio/fio.script; SIZE=4176 # 4176 Mo cumulés par noeud (R/W)
 for N in 1 2 4 8 15 32 64 128; do
    ((NF=$ND+$N-1));
    ((SIZET=$SIZE*$N))
    date +"%s">date.out;
    xdsh hpc-n[$ND-$NF] -f $N "mkdir /$DEST/fio/\$(hostname); cd /$DEST/fio/\$(hostname);
   fio $SCRIPT; mv /$DEST/fio/\$(hostname) /$DEST/fio/\$(hostname).`date +'%Y%m%d-%H%M
   %S'`";
    # Mettre à jour le commentaire dans le printf :
    cat date.out | awk '{t='`date +"%s"`'-$1; printf "%s;8;9000;%s;%s;%s;%s;%s;# 16
   storage/client (409-424) sur edge11\n" , "'$DEST'" , '$N' ,  t , '$SIZE' ,'$SIZET'/t ,
   "'$VARIANTE'" }' >> fio.csv;
 done
```

Procédure de test

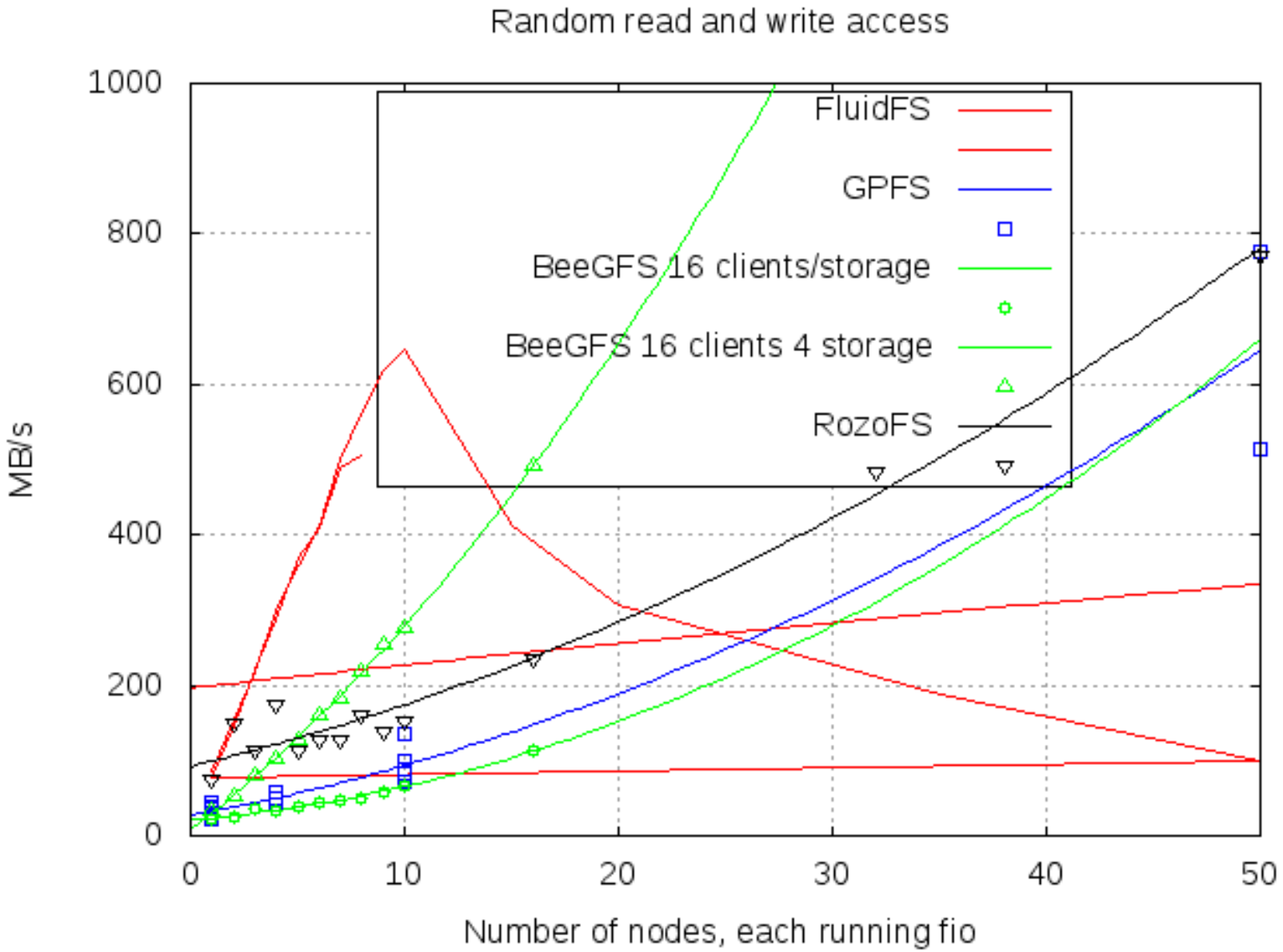
Script fio

```
[global]
 name=bench
 direct=1
 [job1]
 rw=randread
 size=10m
 bs=4k
 [job2]
 rw=randwrite
 size=10m
 bs=4k
 [job3]
 rw=randread
 size=10m
```

…

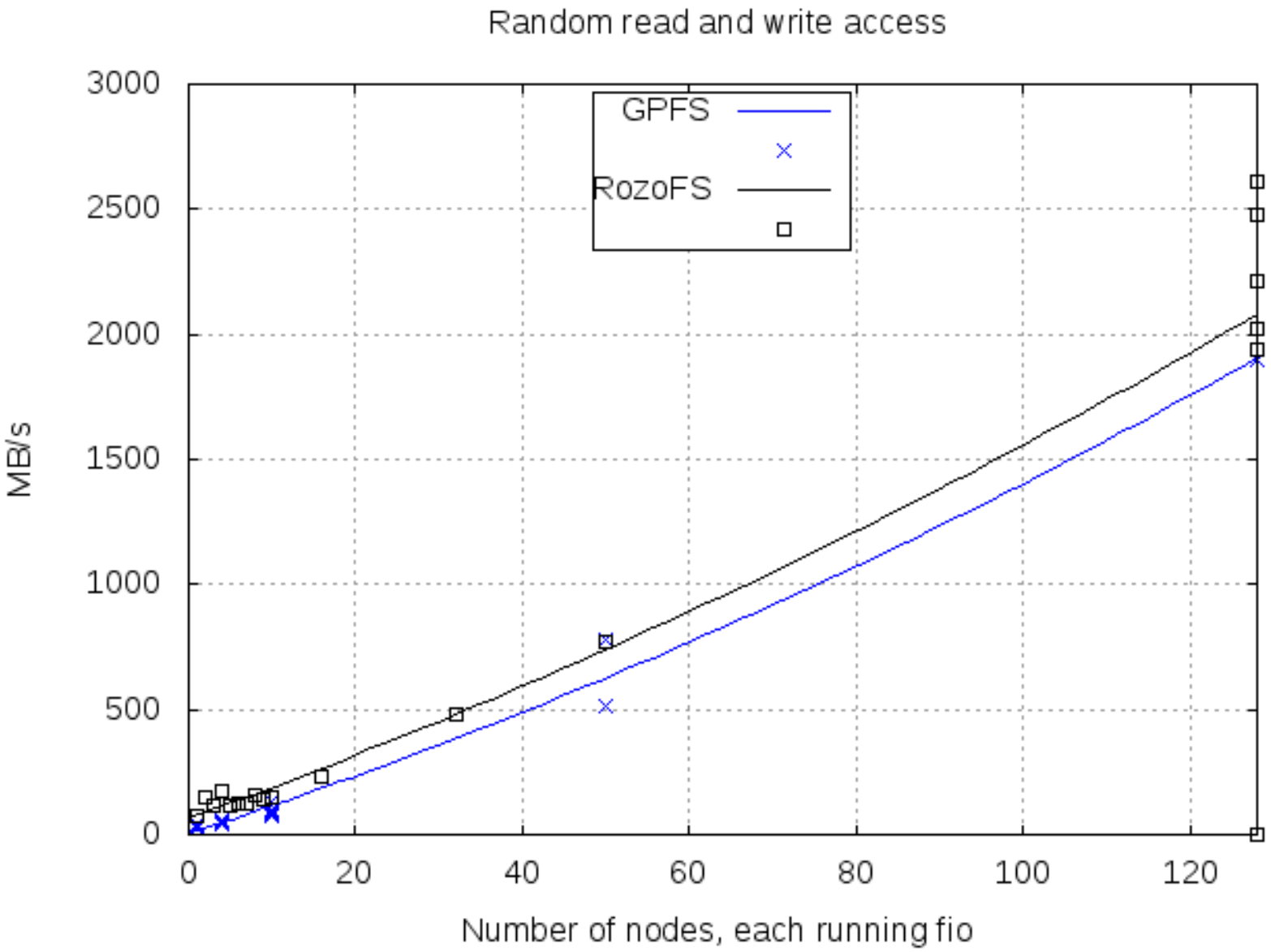
Results

- 1 to 50 nodes



Results

- 1 to 128 nodes



RozoFS specific problems

- Plusieurs problèmes sont survenus sur le POC et sur l'installation en production
  - Certains liés à la faible connectique 1 GbE de nos noeuds
  - D'autres liés spécifiquement à l'environnement HPC : accès simultanés massifs, environnement POSIX et multi-utilisateur
- Très bonne expertise et réactivité de Rozo systems pour résoudre les problèmes
- Par ailleurs : adaptation rapide à nos switchs Juniper

RozoFS specific problems

- Faible connectique 1 GbE des noeuds => un seul storcli et timeouts plus élevés :

1GbE:rozofsnbstorcli=1,rozofsentrytimeout=10,rozofsstoragetimeout=40,rozofsstorclitimeout=60,rozofsexporttimeout=60,rozofsattrtimeout=60,rozofscachemode=2

10GbE:rozofsnbstorcli=2,rozofsentrytimeout=4

- Tuning de la pile IP :

/etc/sysctl.conf

```
net.ipv4.tcp_timestamps      = "0"
net.ipv4.tcp_sack            = "1"
net.ipv4.tcp_low_latency     = "1"
net.ipv4.tcp_adv_win_scale   = "1"
net.core.rmem_default        = "134217728"
net.core.wmem_max            = "134217728"
net.ipv4.tcp_rmem            = "4096 87380 134217728"
net.ipv4.tcp_wmem            = "4096 65536 134217728"
net.core.netdev_max_backlog  = "300000"
net.ipv4.tcp_moderate_rcvbuf = "1"
net.unix.max_dgram_qlen      = "128"
```

UNIVERSITÉ DE STRASBOURG

RozoFS specific problems

Contexte HPC : Accès simultanés massifs, environnement POSIX et multi-utilisateur. Ce contexte est très impactant sur le filesystem.

- 500 noeuds à 2 points de montages ⇒ augmentation du nombre de connexions simultanées supportées par le daemon exportd
- Droits non préservés lors de "cp -p" de GPFS vers RozoFS ⇒ modification du noyau fuse pour support des ACL
- Compilations avec plusieurs threads, faisant des flush simultanés sur le même fichier ⇒ modification du daemon storcli

RozoFS specific problems

- La succession de créations/destructions d'un même répertoire entre plusieurs noeuds à un rythme élevé entraîne aléatoirement des erreurs au mkdir dues à une désynchronisation des noeuds ⇒ timer de cache diminué, augmentation de la fréquence de lookup, factorisation des lookups sur le client pour décharger le serveur de méta-données
- Mauvaise gestion des uid/gid sur les liens symboliques ⇒ corrigé

RozoFS specific problems

- Migration des utilisateurs en cours, bugs levés au fur et à mesure
- Les utilisateurs déjà migrés constatent une amélioration de la performance

- Mise à jour entre le 4 et le 26 mai 2016 :
  - Fichiers vides lors de l'utlisation de routines de type MPI_File_set_view dans des jobs lancés en batch
    ⇒ corrigé par la suppression de certains timers

Plan

- HPC in Strasbourg University
- « My simulation is slow »
- Cascade effect
- Conclusion

Conclusion

- A single application infulenced the whole system
  - Regional computing centers are adaptive !
- *scratch* filesystems are the perfect sandbox
- Data needs to be close to the compute... during the compute !
- SDS, SSD and 7200 RPM disks are the keys to scale-up (capacity) and scale-out (performance) storage, at reasonable prices
- 10GbE and 40GbE is a game changer... makes parallel I/O possible on Ethernet
- We now use 2 SDS systems :

  BeeGFS and RozoFS