BeeGFS

BeeGFS

The Parallel Cluster File System

BeeGFS

The HPC Storage Solution Adopted at the Geneva Observatory

Yves Revaz

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

OBSERVATOIRE DE GENÈVE FONDÉ EN 1772

The Geneva Observatory in a nutshell

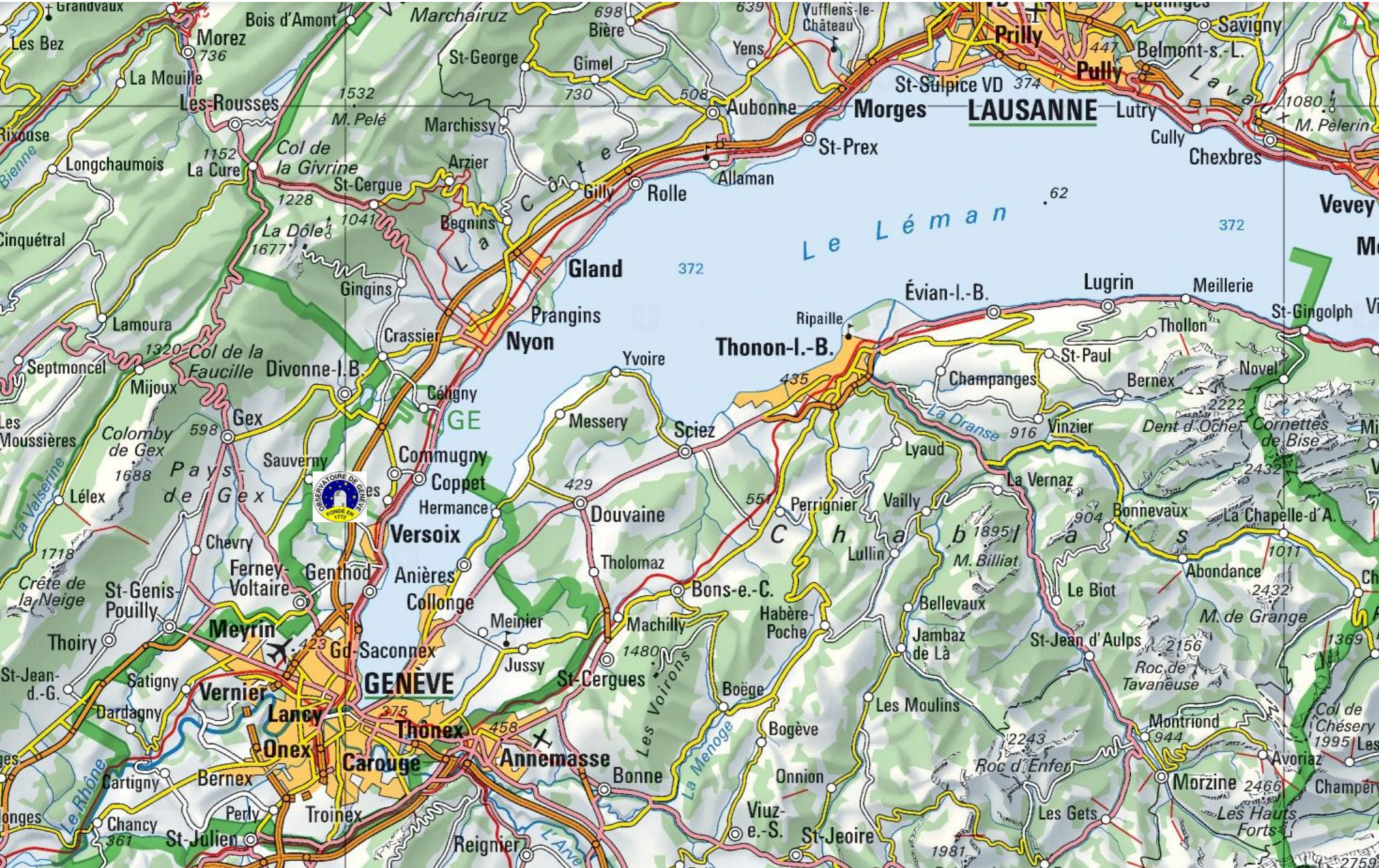
UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

~ 130 collaborators

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

~ 40 collaborators

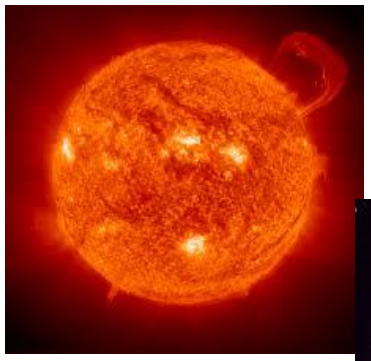
The Geneva Observatory in a nutshell



The Geneva Observatory in a nutshell

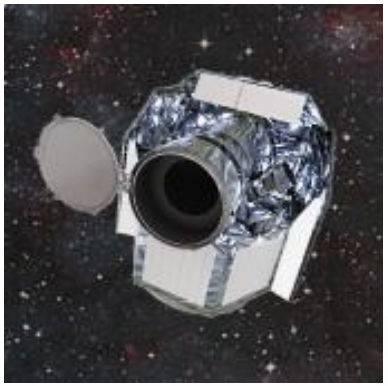
Main fields of research

- Stellar Physics
- Galaxy and Cosmology
- High energy astrophysics
- Extra-solar planets



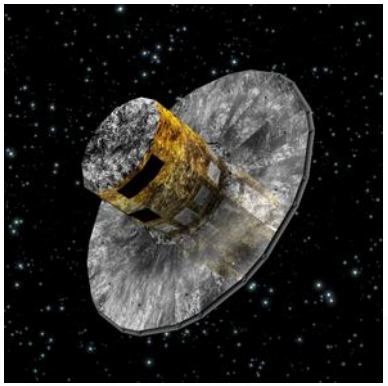
The Geneva Observatory in a nutshell

Tight connections with ESA space missions



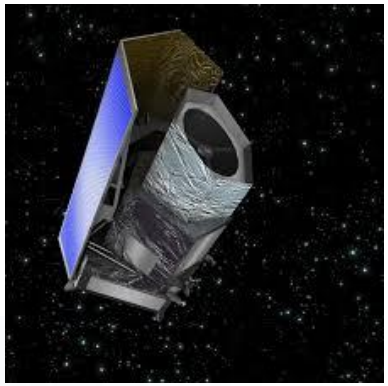
Cheops

2018



Gaia

in opperation



Euclid

2021

Integral

in opperation

HPC support to research and space missions

- 160 kW computing (data) center
- ~ 1500 cores (cluster)
- an IT team of 6 people

The Geneva Observatory HPC Storage System

HPC storage needs and constraints

- HPC file system
  - supporting concurrent access (MPI-IO, HDF5)
  - accessed by about 1500 cores (~5000 in 2022)
- Capacity/Extensibility
  - about 1Po (about 6 Po in 2022)
- Versatility
  - should work both as
    - an efficient HPC storage
    - a long term storage (large amount of data must be regulatrly re-processed)
  - should support a large variety of applications dealing with different files

HPC storage needs and constraints

- Data integrity
  - data redundancy (long term storage)
  - no need for snapshots
- Price
  - the lowest as possible
- No need for object-based storage
- No need for high-availability

HPC storage solutions

- GPFS  too expansive
- Panasas  too expansive
- Luster  too complex
- PVFS  bad experience

BeeGFS  let’s try it !

BeeGFS

The Parallel Cluster File System

- BeeGFS is a truly parallelized cluster file system offering a common storage space, spread over several servers.

- Started in 2005, and initially called FhGFS (Fraunhofer) as it was developed by the Fraunhofer Center for High Performance Computing.

- Now developped by a company ThinkParQ, offering support and consulting.

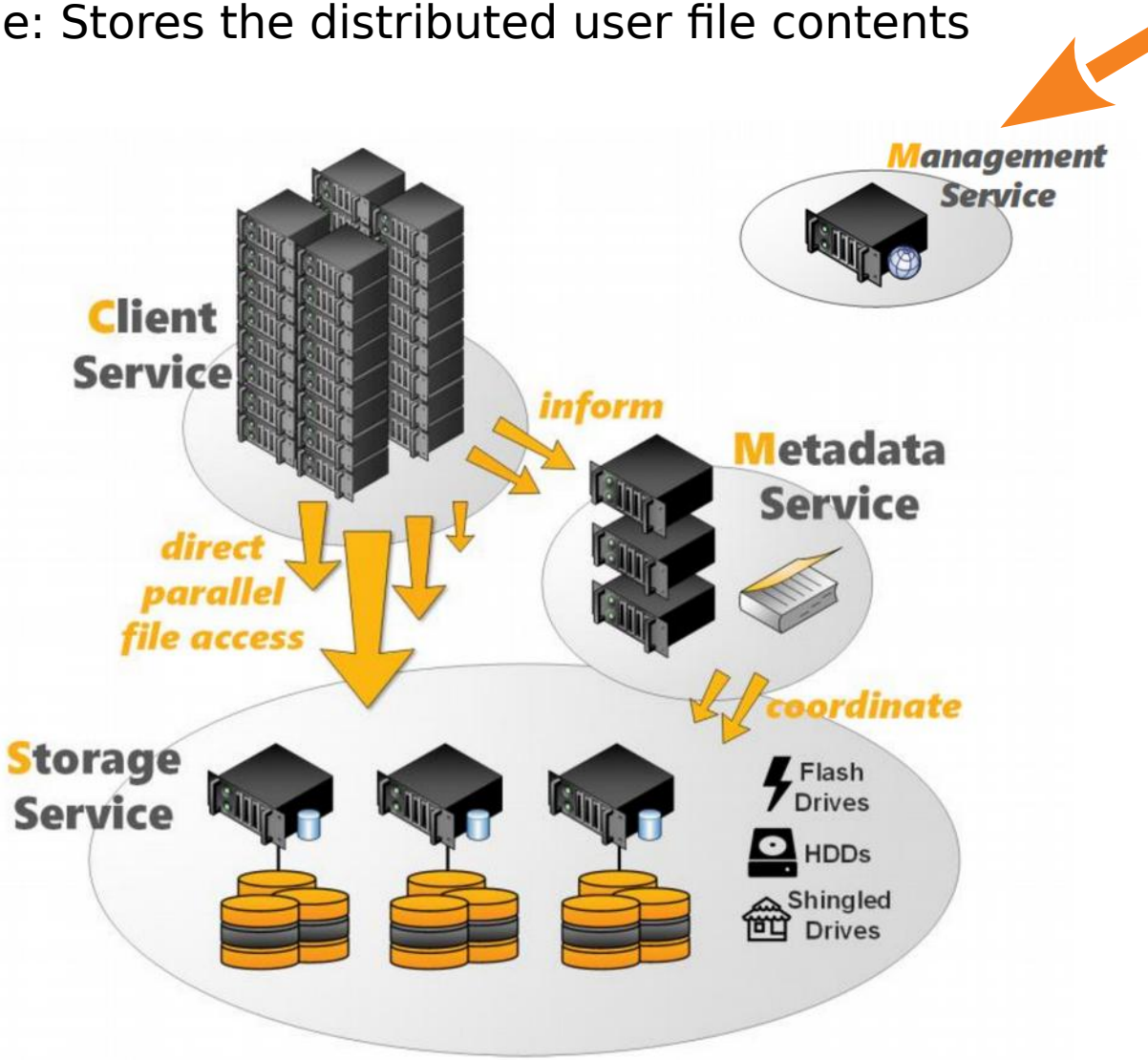
BeeGFS®

The BeeGFS many advantages

- Supported on linux systems
- Hardware independent
- Easy to install (rpms), fast deployment (few configuration needed)
- Easy to manage (few commands)
- Easy to extend
- Offers data redumdancy options
- Support different network types
  - Standard TCP/IP network
  - RDMA-capable networks like InfiniBand (IB) or Omni-Path (OPA)
- Free of charge
  - client: GPLv2 licence
  - Server: BeeGFS licence

The BeeGFS architecture: 4 services

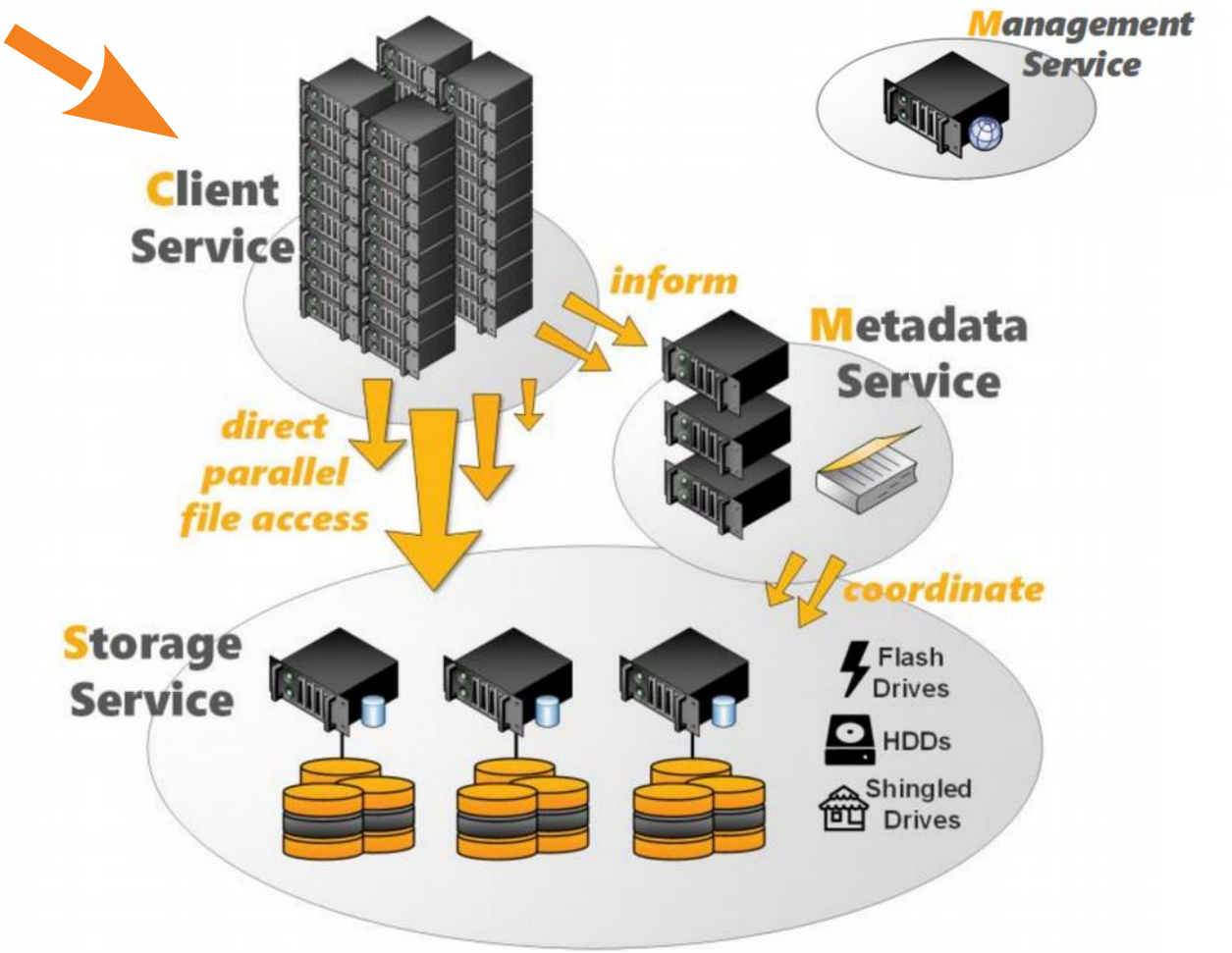
- Management service: A registry and watchdog for all other services
- Client service: Mounts the file system to access the stored data
- Metadata service: Stores access permissions and striping information
- Storage service: Stores the distributed user file contents



BeeGFS®

The BeeGFS architecture: 4 services

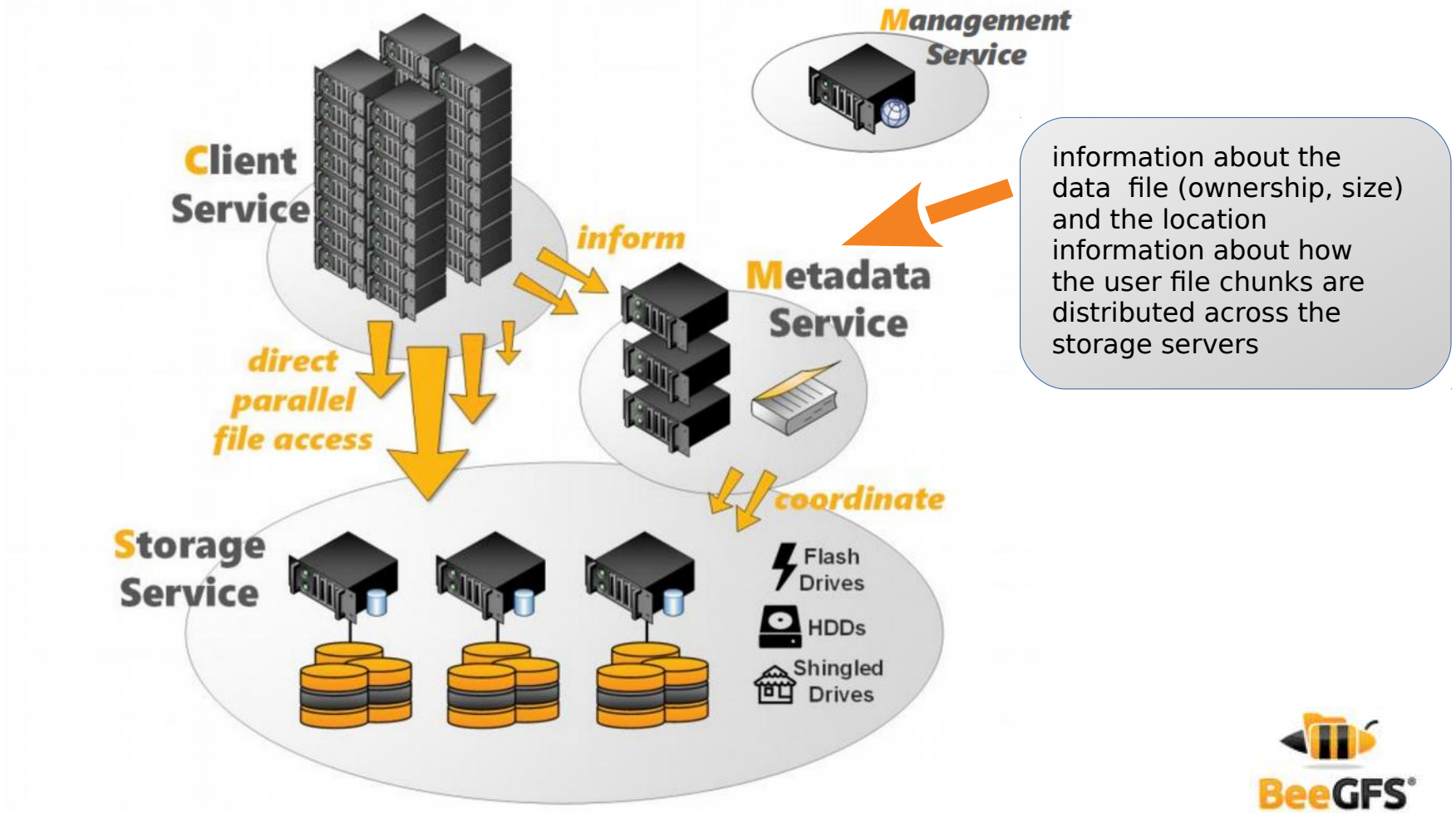
- Management service: A registry and watchdog for all other services
- Client service: Mounts the file system to access the stored data
- Metadata service: Stores access permissions and striping information
- Storage service: Stores the distributed user file contents



BeeGFS®

The BeeGFS architecture: 4 services

- Management service: A registry and watchdog for all other services
- Client service: Mounts the file system to access the stored data
- Metadata service: Stores access permissions and striping information
- Storage service: Stores the distributed user file contents

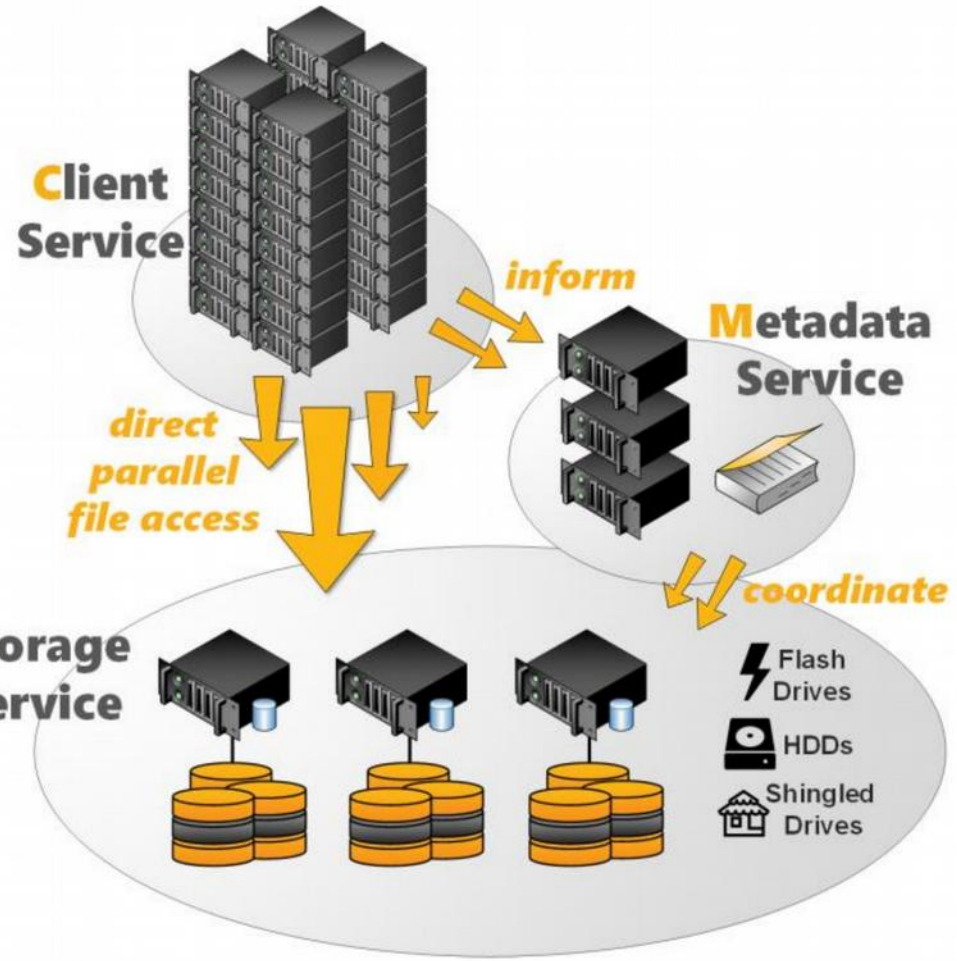


BeeGFS®

The BeeGFS architecture: 4 services

- Management service: A registry and watchdog for all other services
- Client service: Mounts the file system to access the stored data
- Metadata service: Stores access permissions and striping information
- Storage service: Stores the distributed user file contents

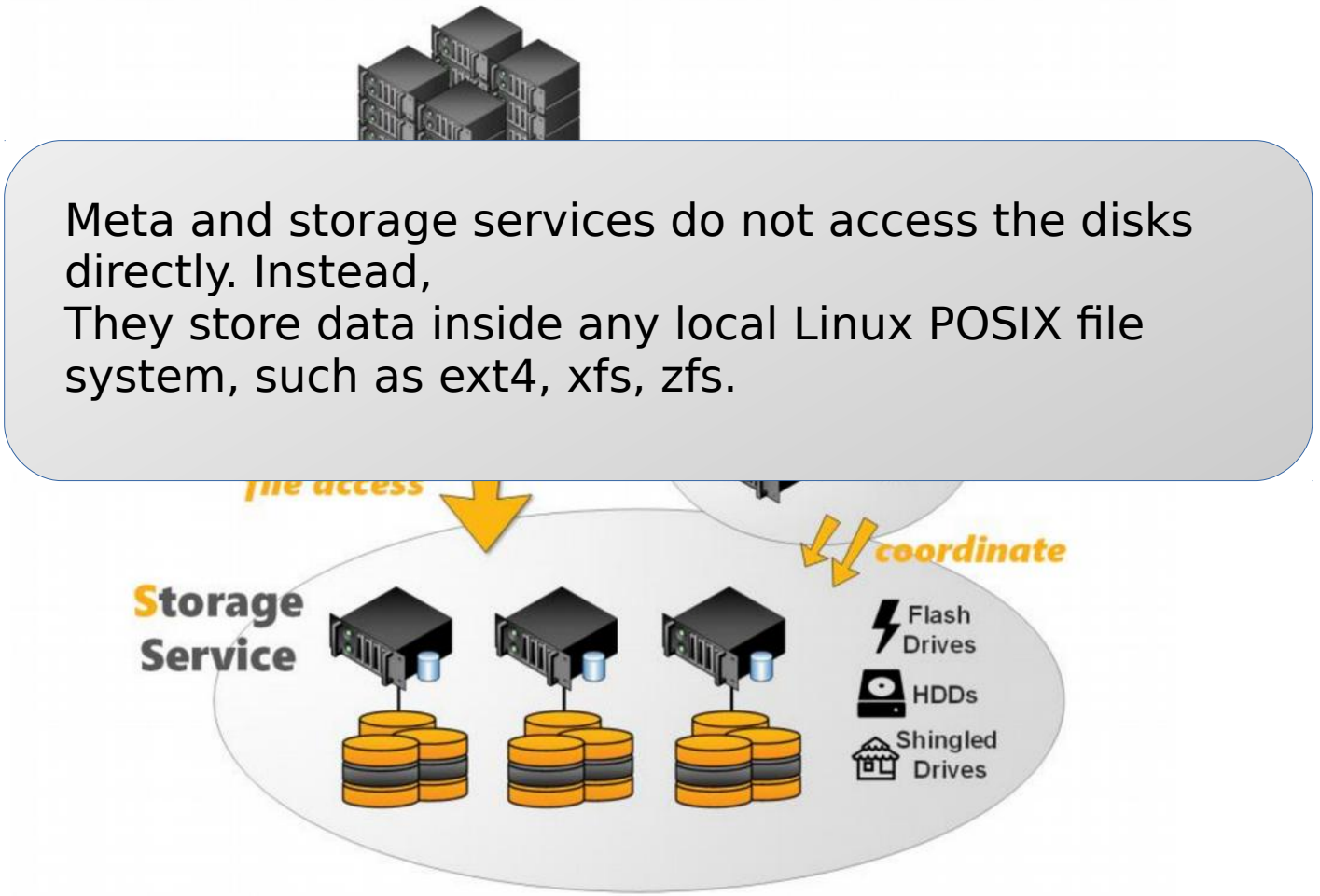
store striped user file contents one servers may have one or multiple storage targets.



BeeGFS®

The BeeGFS architecture: 4 services

- Management service: A registry and watchdog for all other services
- Client service: Mounts the file system to access the stored data
- Metadata service: Stores access permissions and striping information
- Storage service: Stores the distributed user file contents

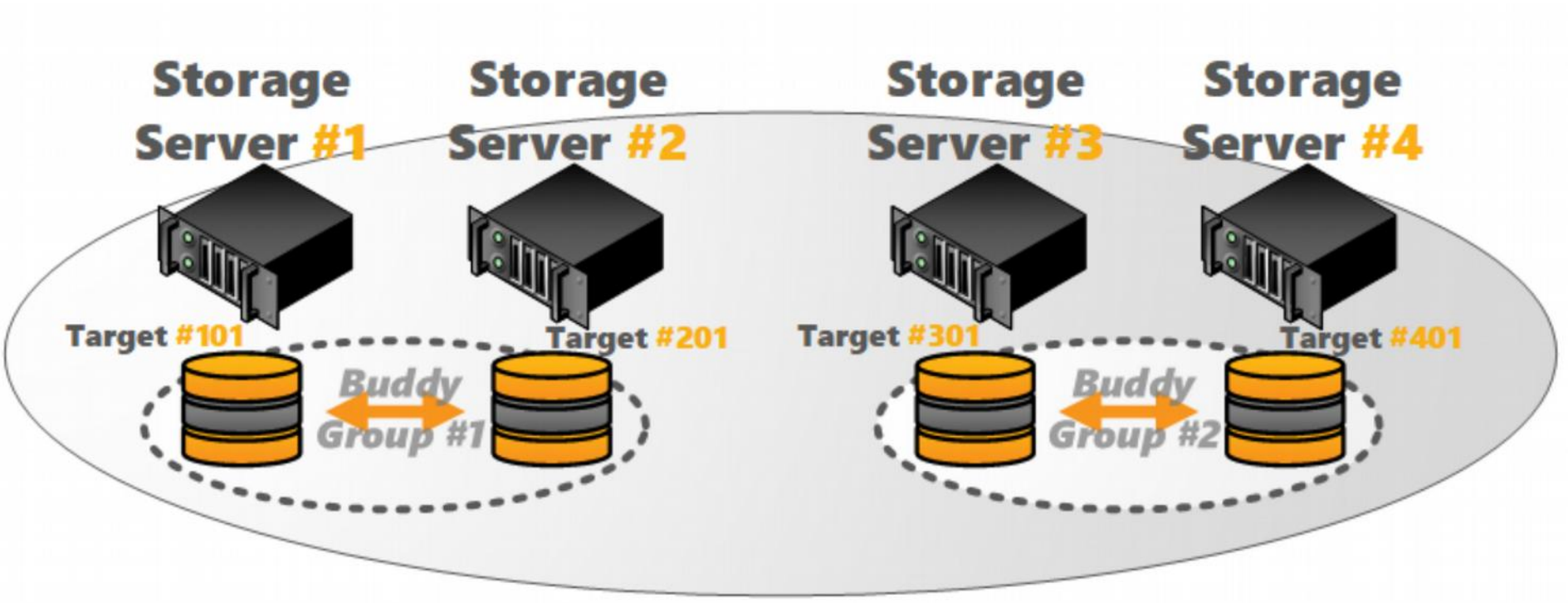


Meta and storage services do not access the disks directly. Instead,
They store data inside any local Linux POSIX file system, such as ext4, xfs, zfs.

BeeGFS®

Security : meta-data or data mirroring

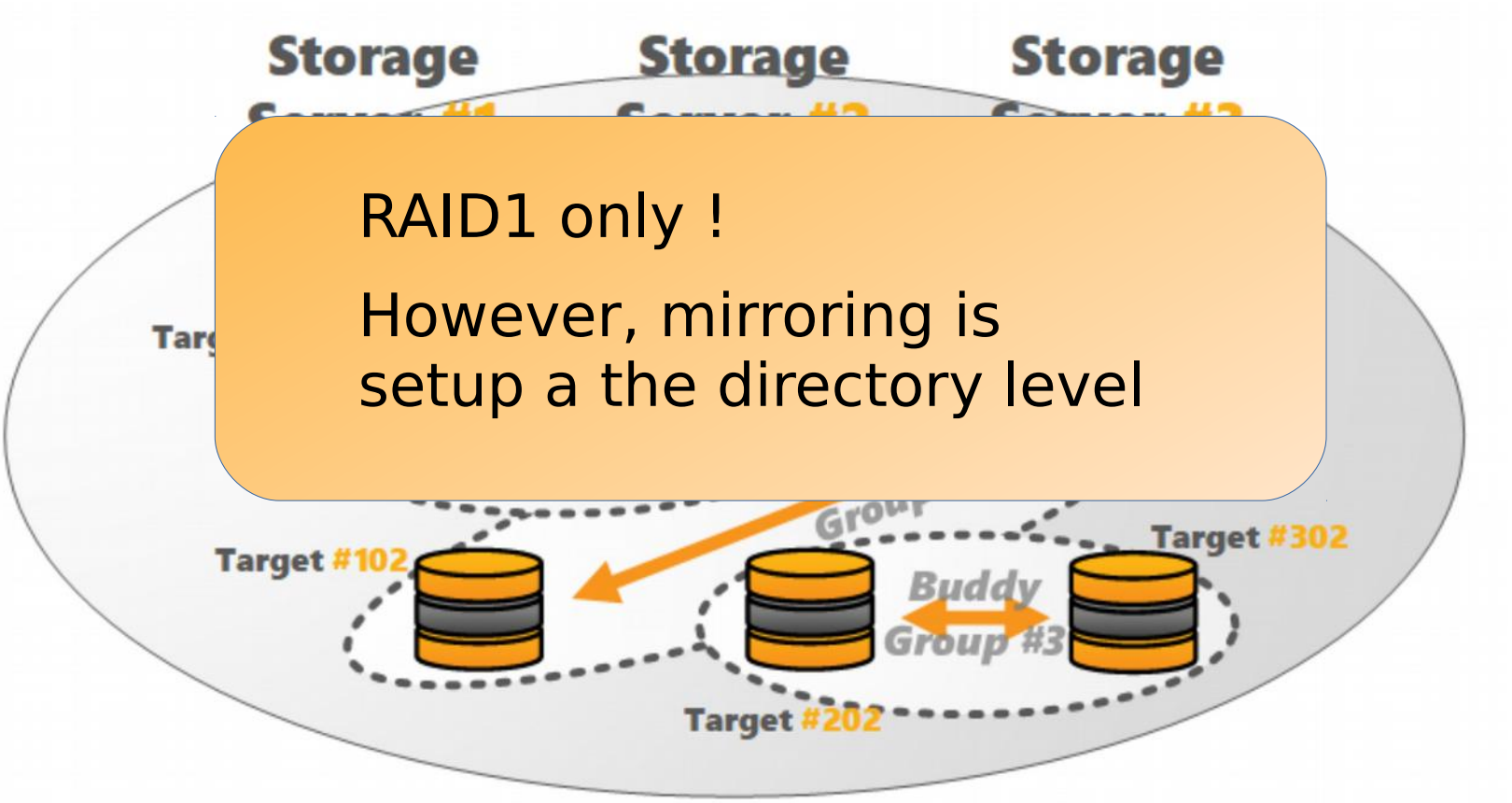
- Buddy groups : groups of targets (meta-data or data)



BeeGFS®

Security : meta-data or data mirroring

- Buddy groups : groups of targets (meta-data or data)



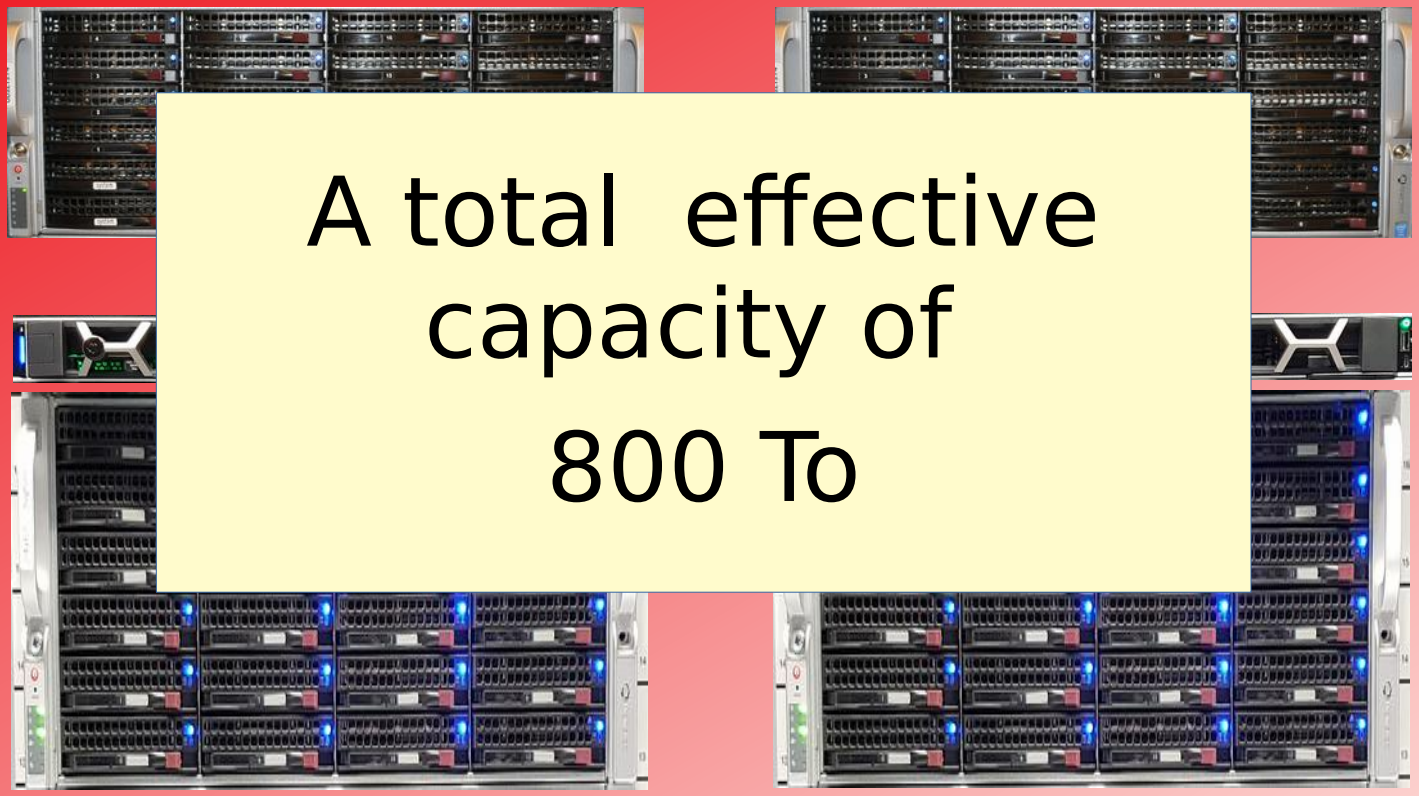
BeeGFS®

Details of the hardware of our BeeGFS system solution

Overview of the BeeGFS system

2 meta-data servers

4 data servers



A total effective capacity of

800 To

Interconnect : Infiniband 40 Gb/sec (4X QDR)

Details of the components

2 meta-data servers

Server Type	Transtec Calleo 2280s
CPUs	2x Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz 16 cores
RAM	64 Go
Disks	4x 900Go SSD (RAID5 3+1)
Partitions	/dev/sdb RAID5 2.7TB

Details of the components

2 data servers

Server Type	Transtec Calleo 4280
CPUs	2x Intel(R) Xeon(R) CPU E5-2609 v4 @ 1.70GHz 16 cores
RAM	128 Go
Disks	24x 7.2To HDD
Partitions	/dev/md0 (RAID 6 10+2) 72 TB /dev/md1 (RAID 6 10+2) 72 TB

Details of the components:

2 data servers

Server Type	Dell R630 + 2x JBOD
CPUs	2x Intel(R) Xeon(R) Gold 5122 CPU @ 3.60GHz 8 cores
RAM	128 Go
Disks	24x 5.5To HDD 24x 7.2To HDD
Partitions	/dev/md0 (RAID 6 10+2) 55 TB /dev/md1 (RAID 6 10+2) 55 TB /dev/md2 (RAID 6 10+2) 72 TB /dev/md3 (RAID 6 10+2) 72 TB

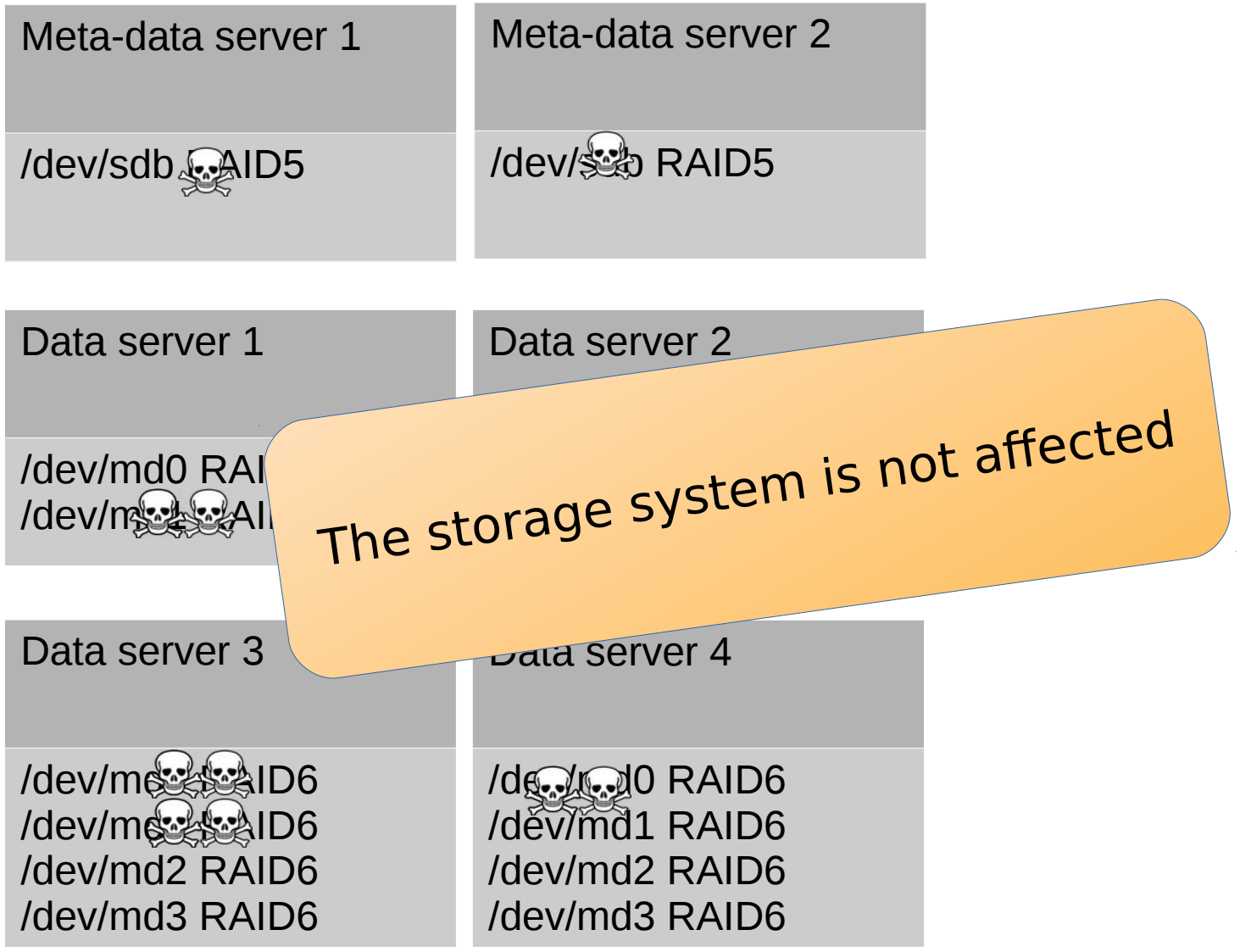
Data redundancy : disk RAIDS

Meta-data server 1	Meta-data server 2
/dev/sdb RAID5	/dev/sdb RAID5

Data server 1	Data server 2
/dev/md0 RAID6 /dev/md1 RAID6	/dev/md0 RAID6 /dev/md1 RAID6

Data server 3	Data server 4
/dev/md0 RAID6 /dev/md1 RAID6 /dev/md2 RAID6 /dev/md3 RAID6	/dev/md0 RAID6 /dev/md1 RAID6 /dev/md2 RAID6 /dev/md3 RAID6

Data redundancy : disk RAIDS

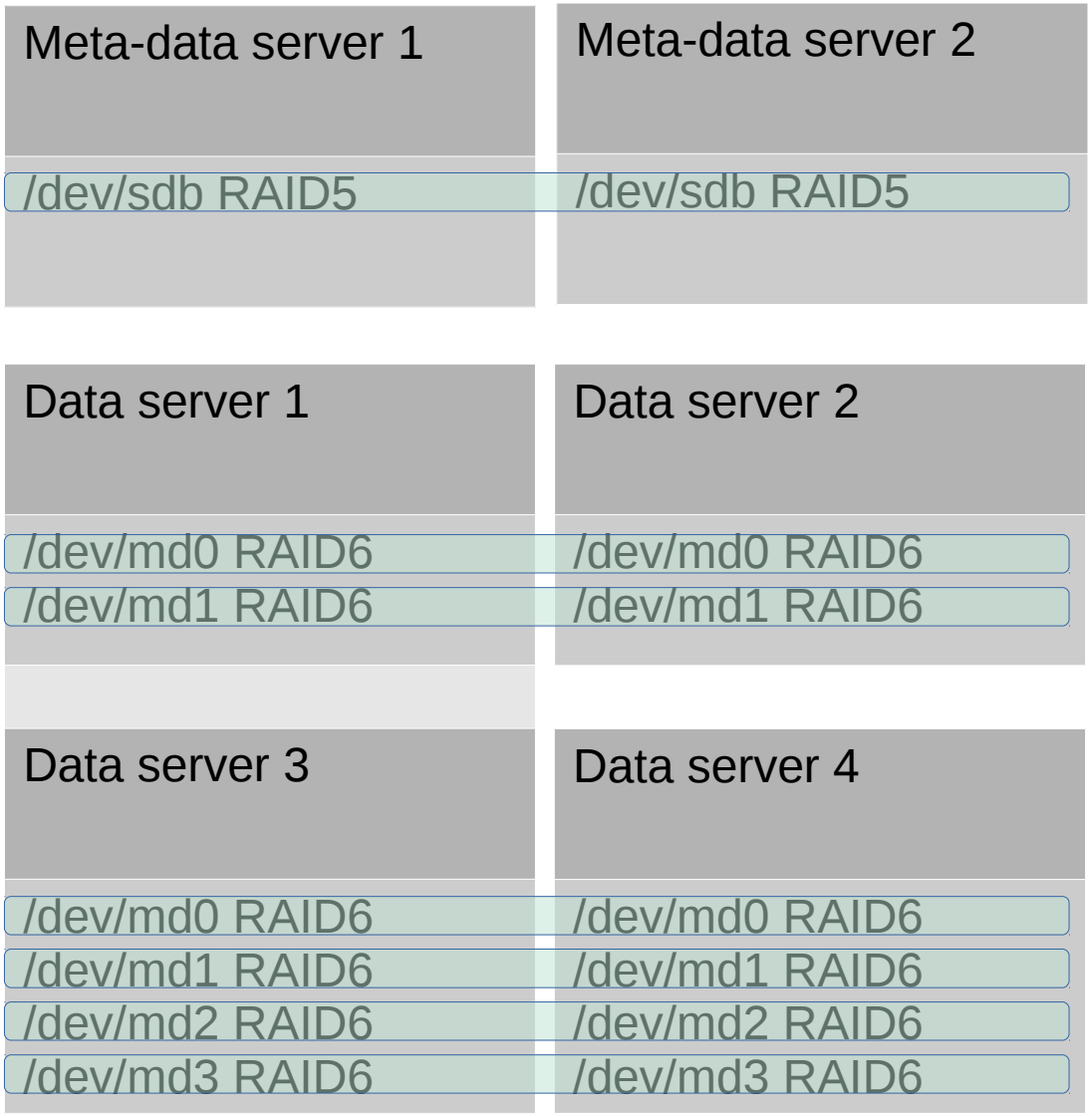


Data redundancy : disk RAIDS

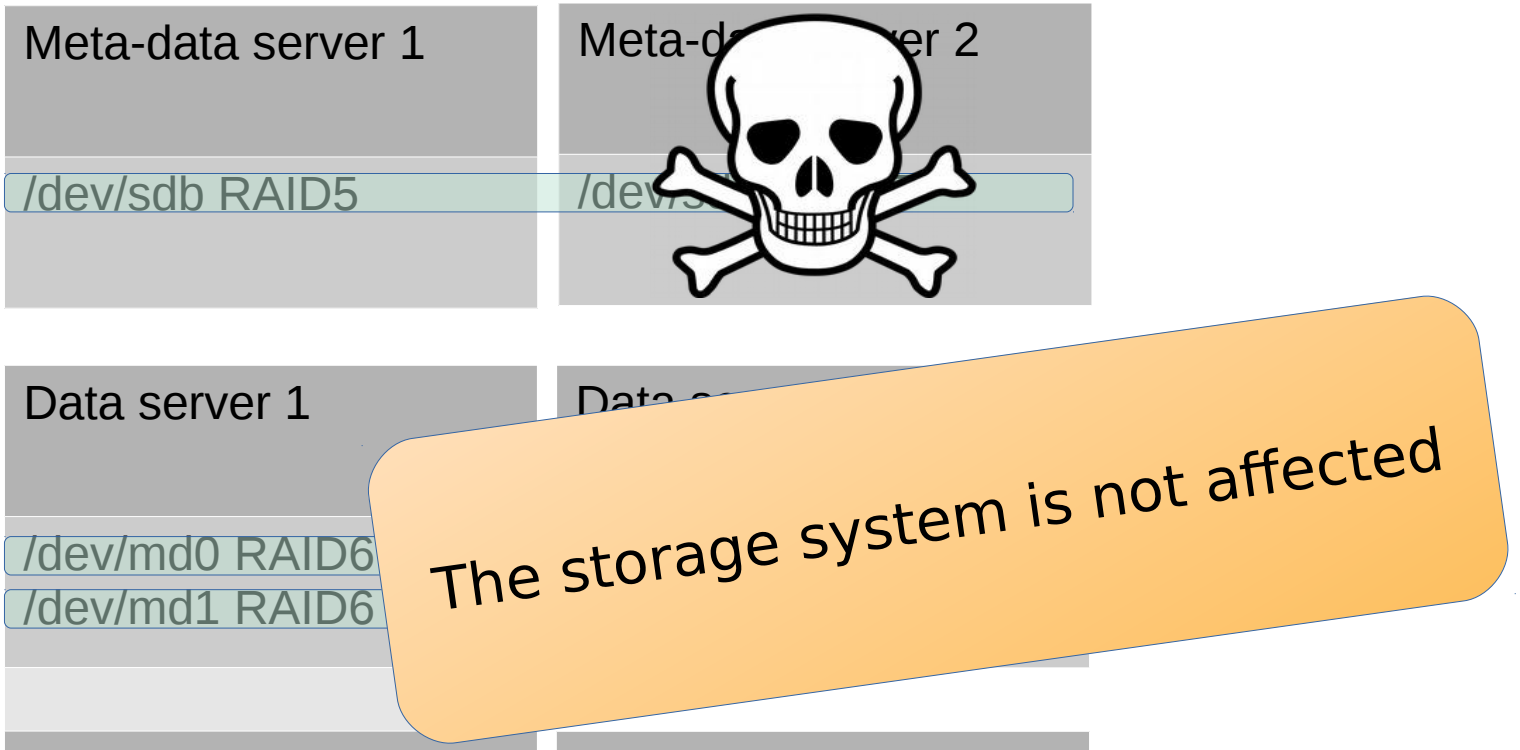
Meta-data server 1	Meta-data server 2
/dev/sdb RAID5	/dev/sdb RAID5
/dev/md1 RAID6	/dev/md0 RAID6 /dev/md1 RAID6
Data server 3	Data se
/dev/md0 RAID6 /dev/md1 RAID6 /dev/md2 RAID6 /dev/md3 RAID6	/dev/md0 RAID6 /dev/md1 RAID6 /dev/md2 RAID6 /dev/md3 RAID6

data integrity is conserved
after the re-installation
of the data server

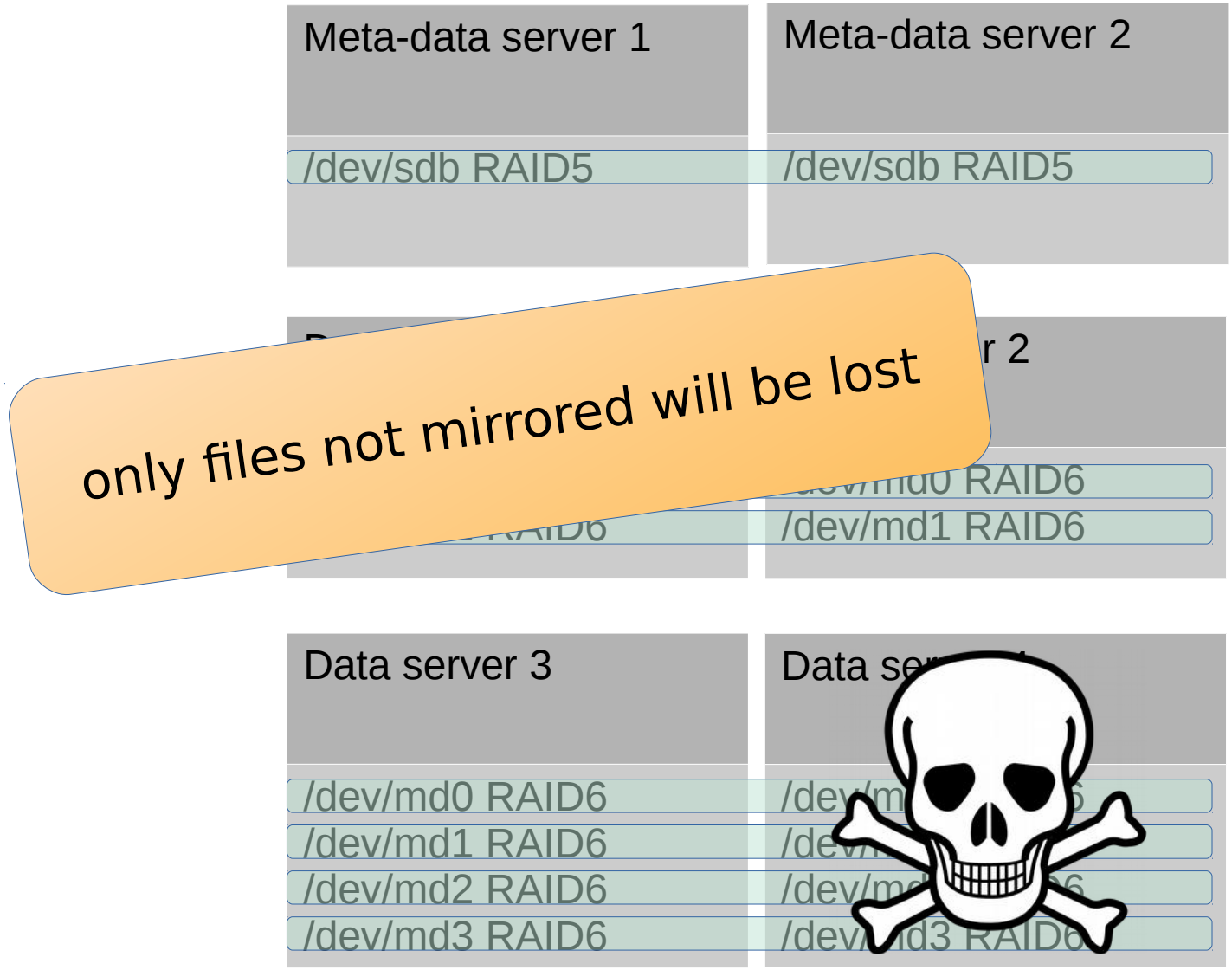
Data redundancy : buddy mirrors



Data redundancy : buddy mirrors



Data redundancy : buddy mirrors



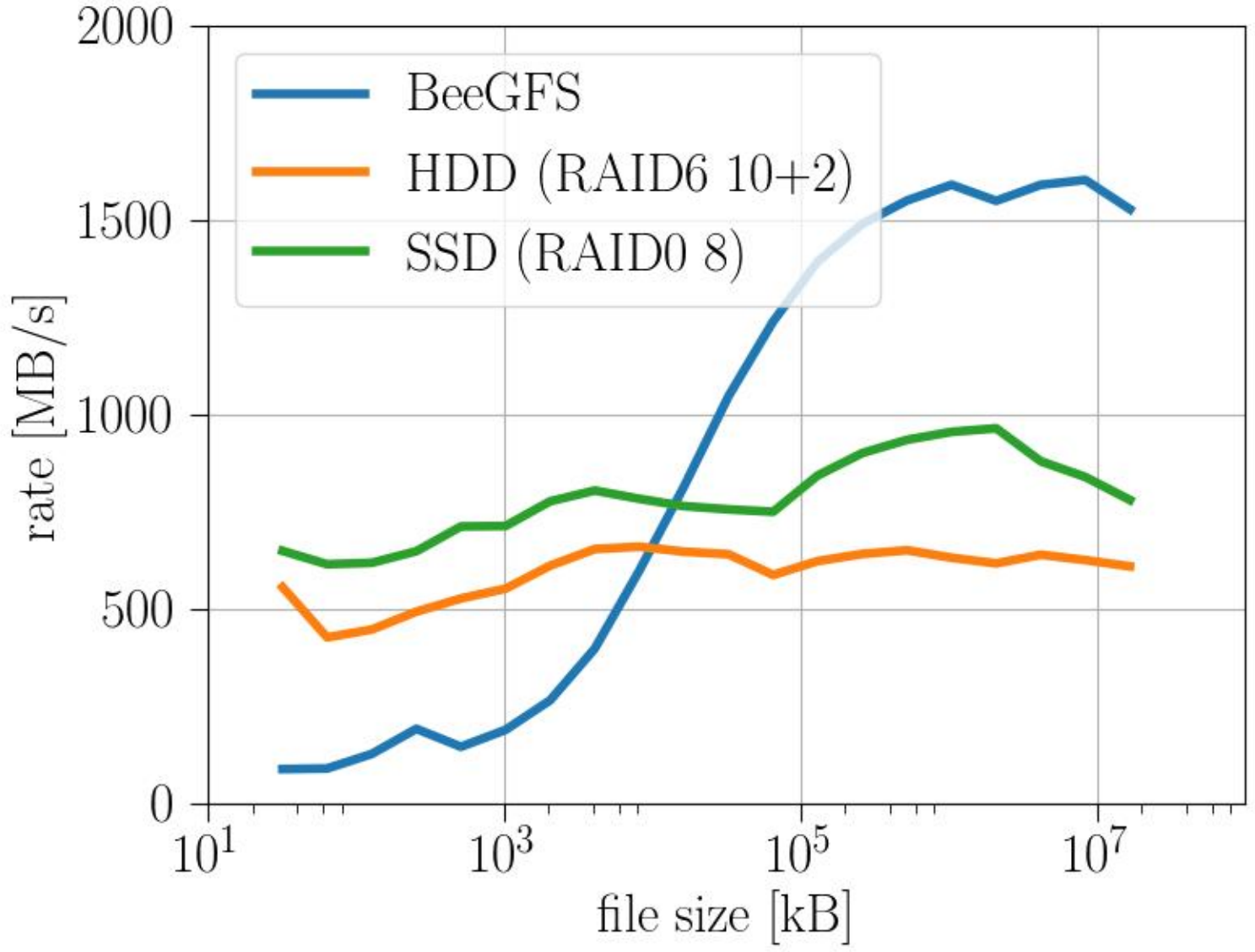
Benchmarks

Default setup:

- chunk size : 1Mo
- number of targets : 12

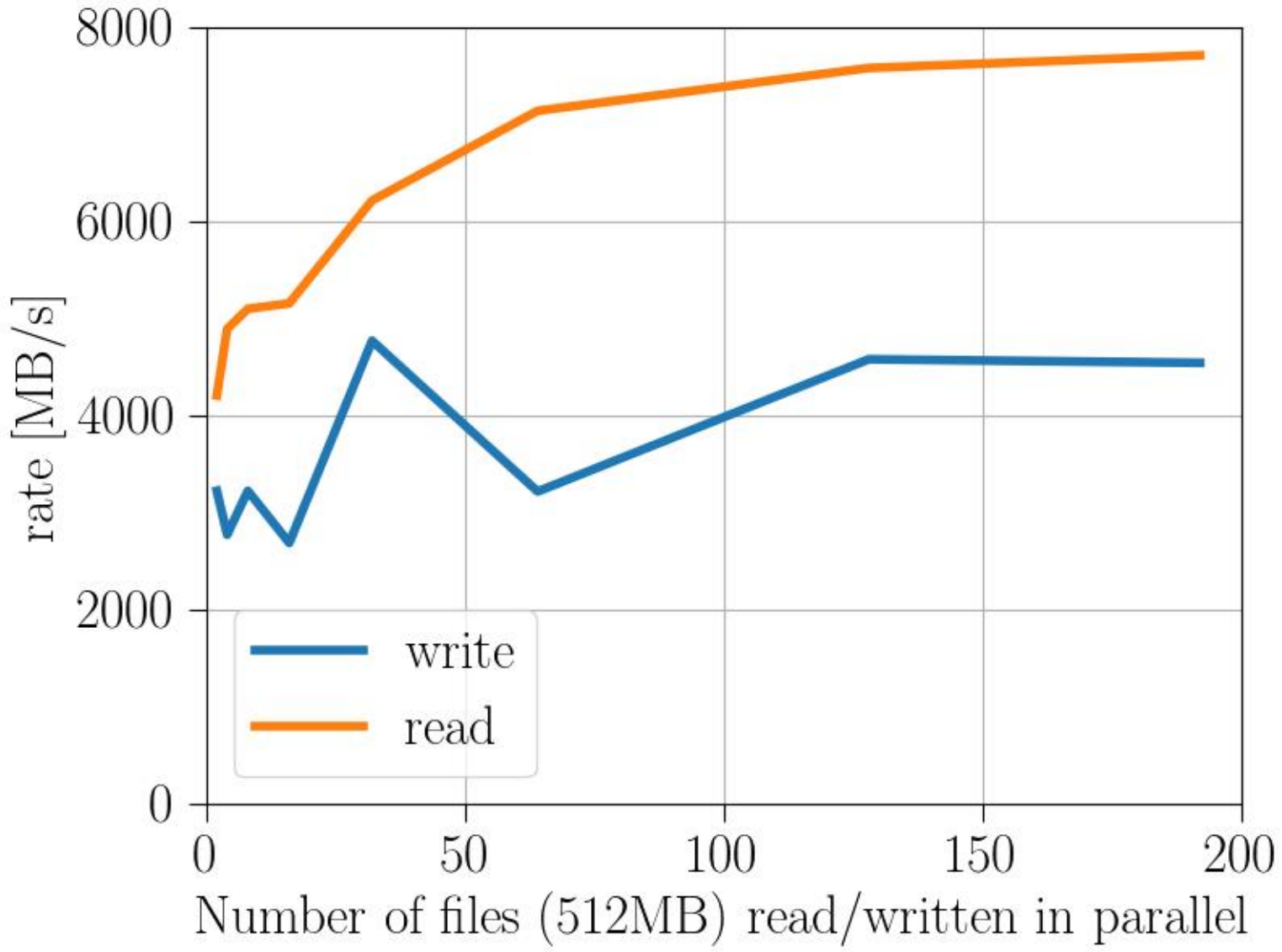
Benchmarks I

1 task writes one file



Benchmarks II : iobench

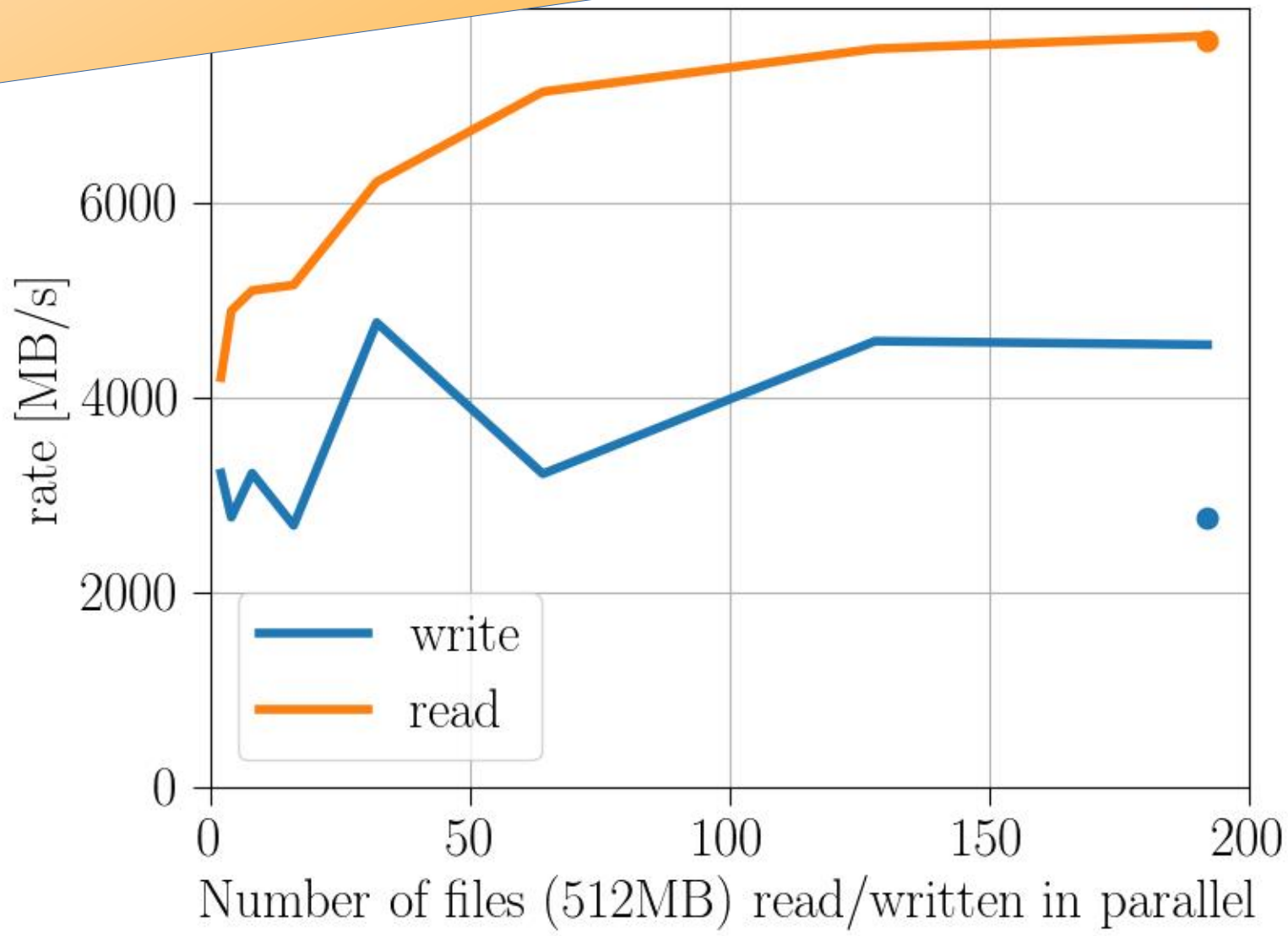
N tasks read/write each at the same time one 512MB file



Benchmarks II : iobench

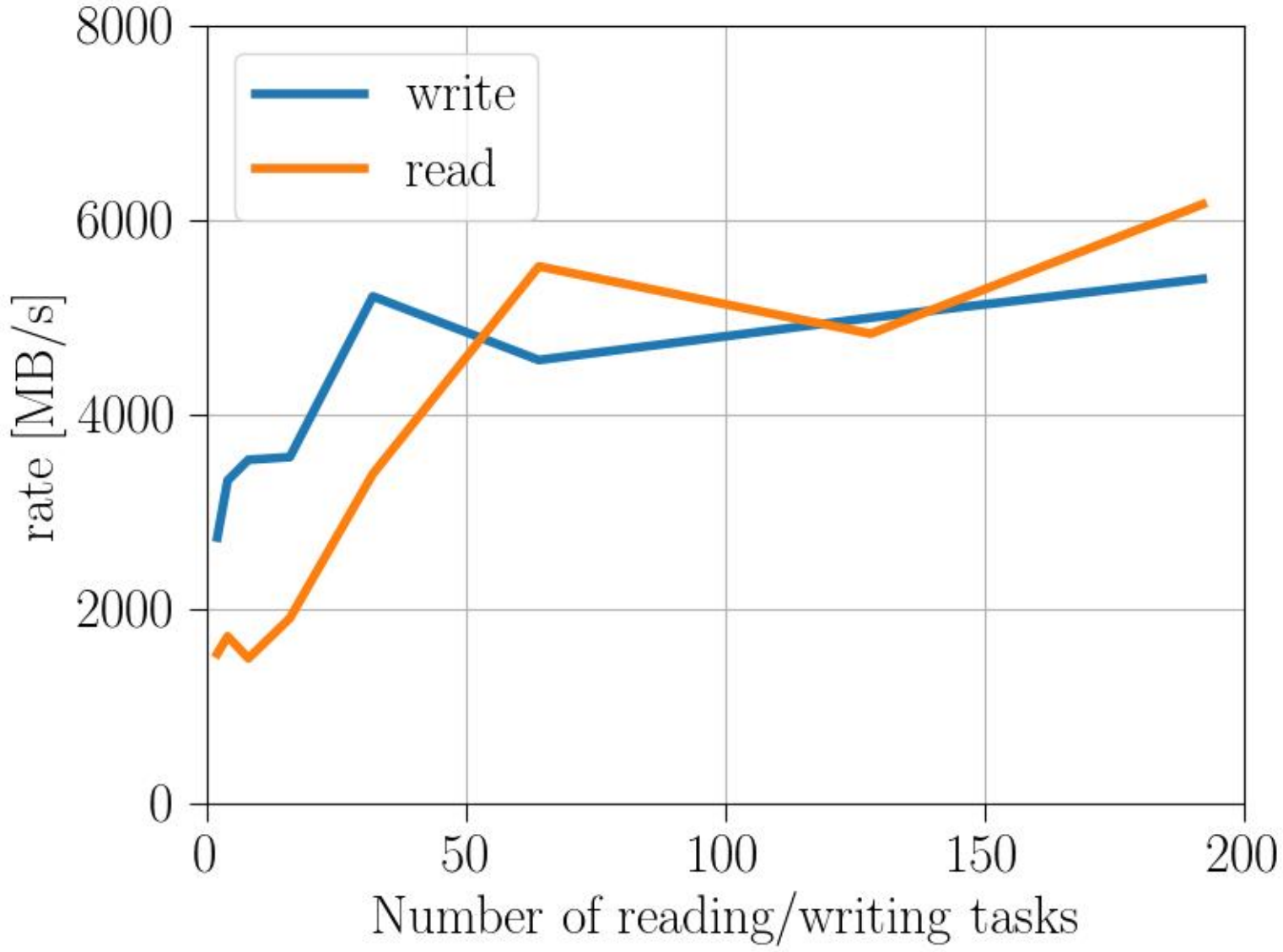
only one storage target per file

me time one 512MB file



Benchmarks II : iobench

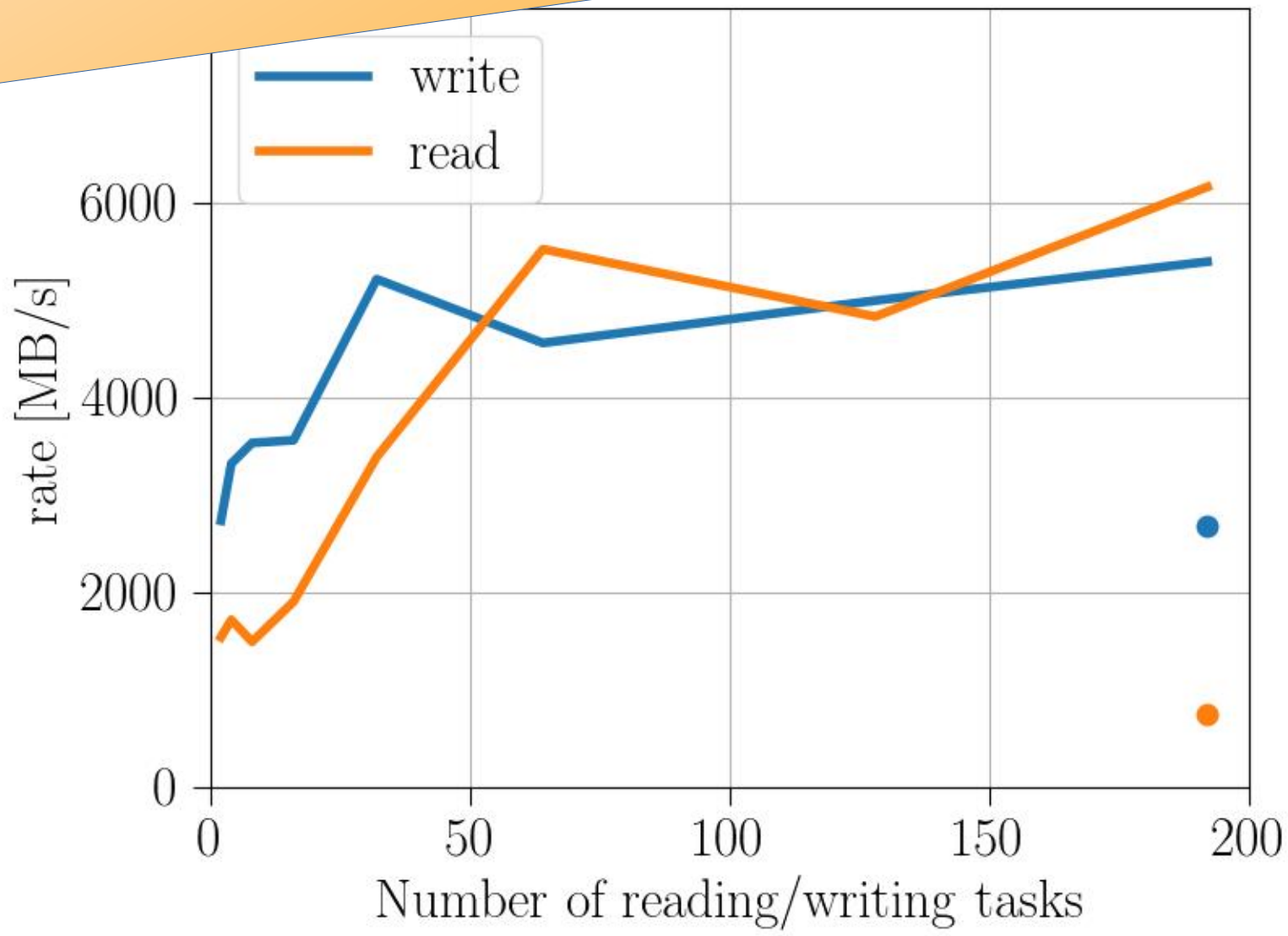
N tasks read/write each 512MB in the same file



Benchmarks II : iobench

only one storage target per file

n the same file



Conclusions

- We are globally satified with the BeeGFS solution
  - Easy deployment/installation/management/extension
  - Performances that fulfil our requirements

- Improvements
  - Only RAID1 is supported for data mirroring. A RAID5 will be welcome in the future.
  - User/Groups quota are not optimal. Provided only globally, but not at a directory level.

The End