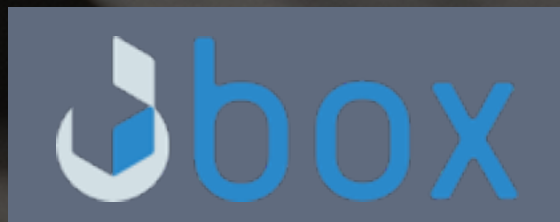
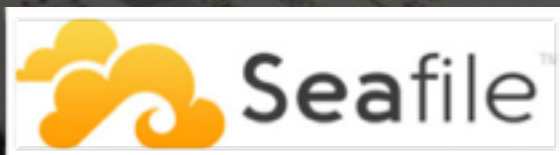


# SYNC & SHARE TO REPLACE HPC HOMES?



Krzysztof Wadówka,  
Maciej Brzeźniak,  
Marcin Pospieszny,  
Piotr Brona,  
Radosław Januszewski

HPC Department





# AGENDA

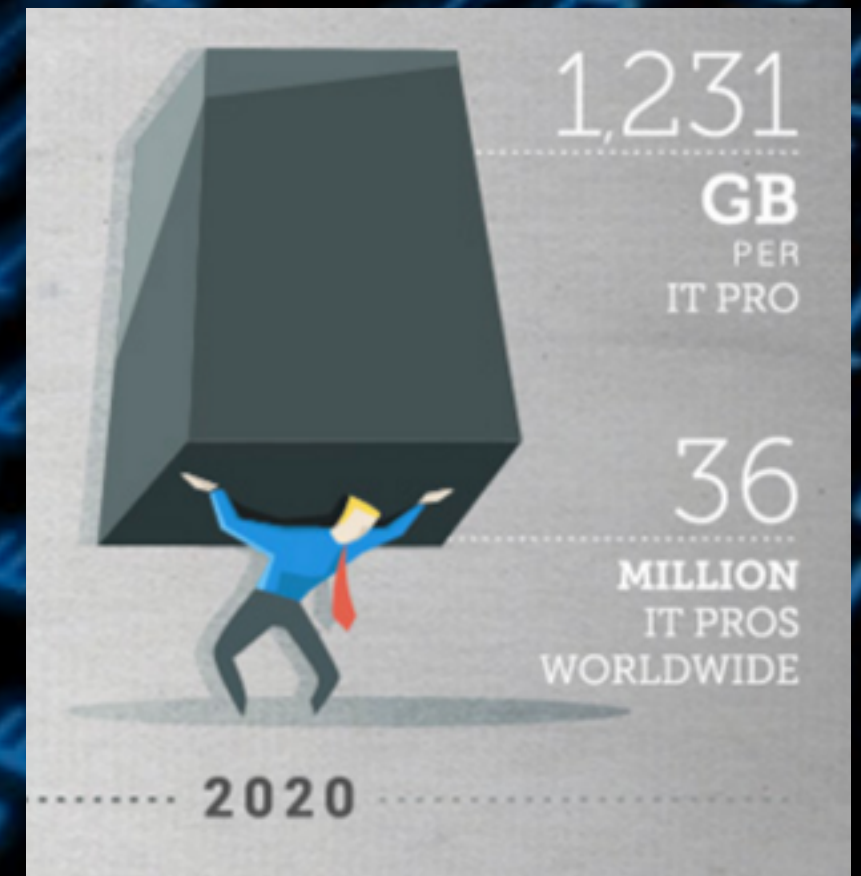


- Challenges / background
- Why Seafile?
- HPC @PSNC & PIONIER-BOX: who we are, aims
- Our setup & tests
- Observations & discussion



# DATA MGMT IN HPC

- **DATA VOLUMES:**
  - PetaBytes...
  - 10-100s **GB/s**, 100s k **IOPS**
- **HPC storage:**
  - fast but hermetic
  - not easy to access
- **Dilemma: performance vs usability:**
  - SFTP, GridFTP, NFS - no longer a solution
  - HPC storage especially hermetic
- **Users:**
  - want **performance of Lustre**...
  - with the Dropbox' **ease of use**



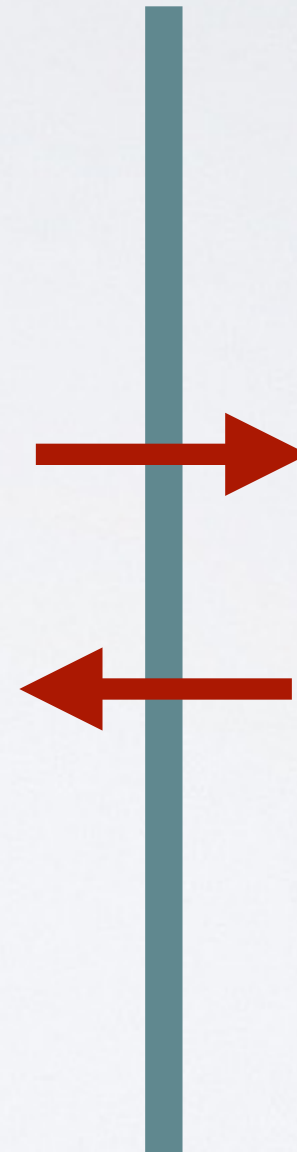
Source: IDC



# TYPICAL DATA FLOW IN HPC



User & data



**Barrier**



Storage

# TARGET DATA FLOW IN HPC



User & data



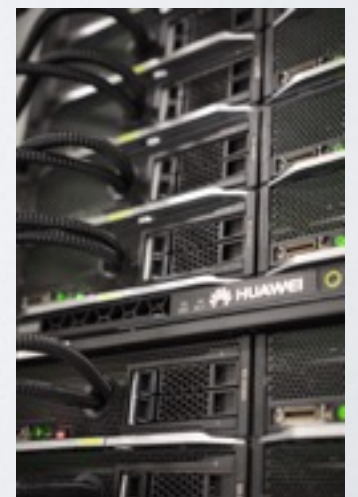
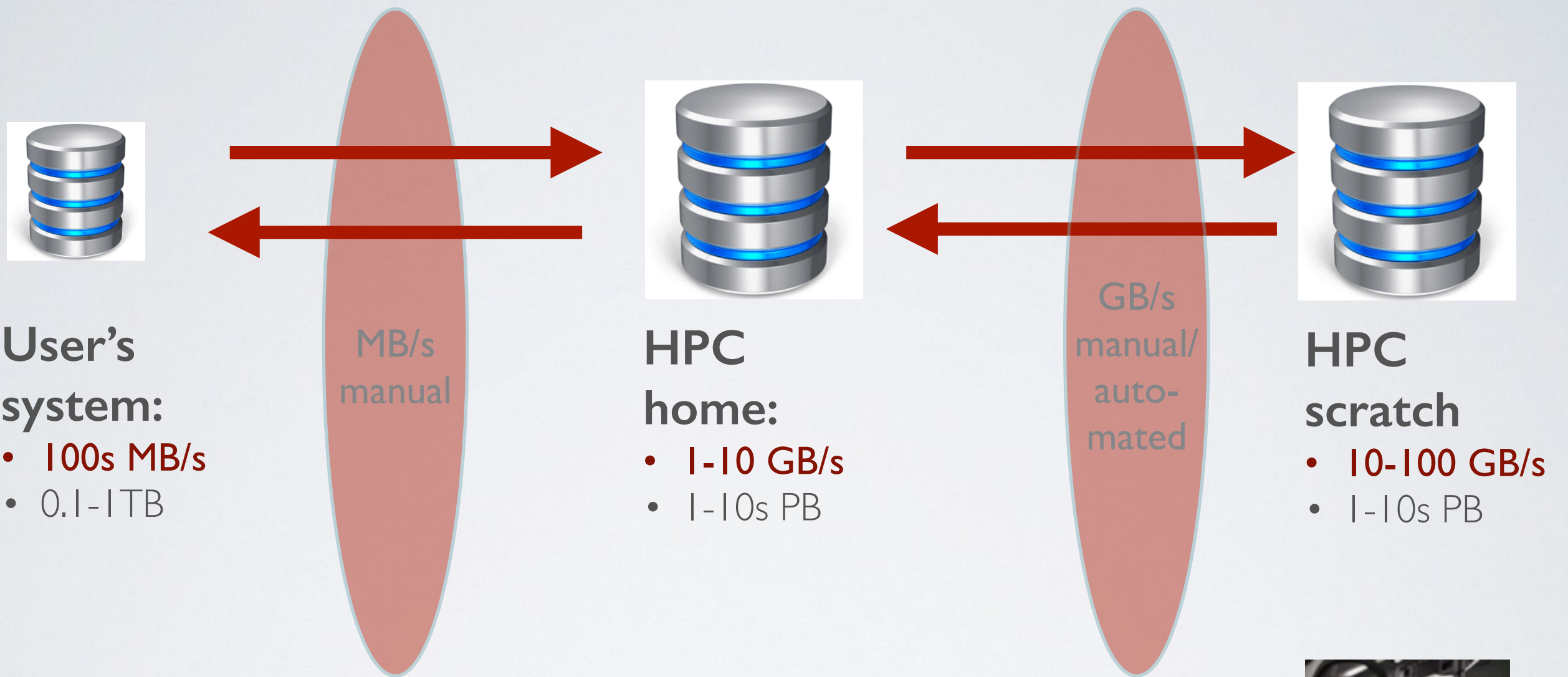
**Interface**



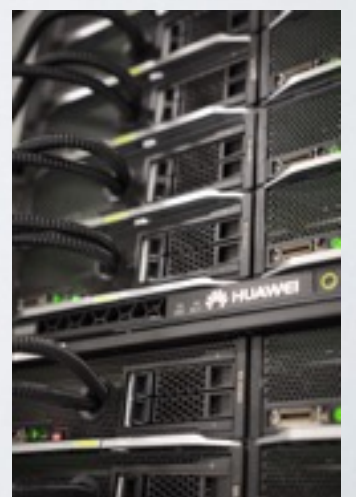
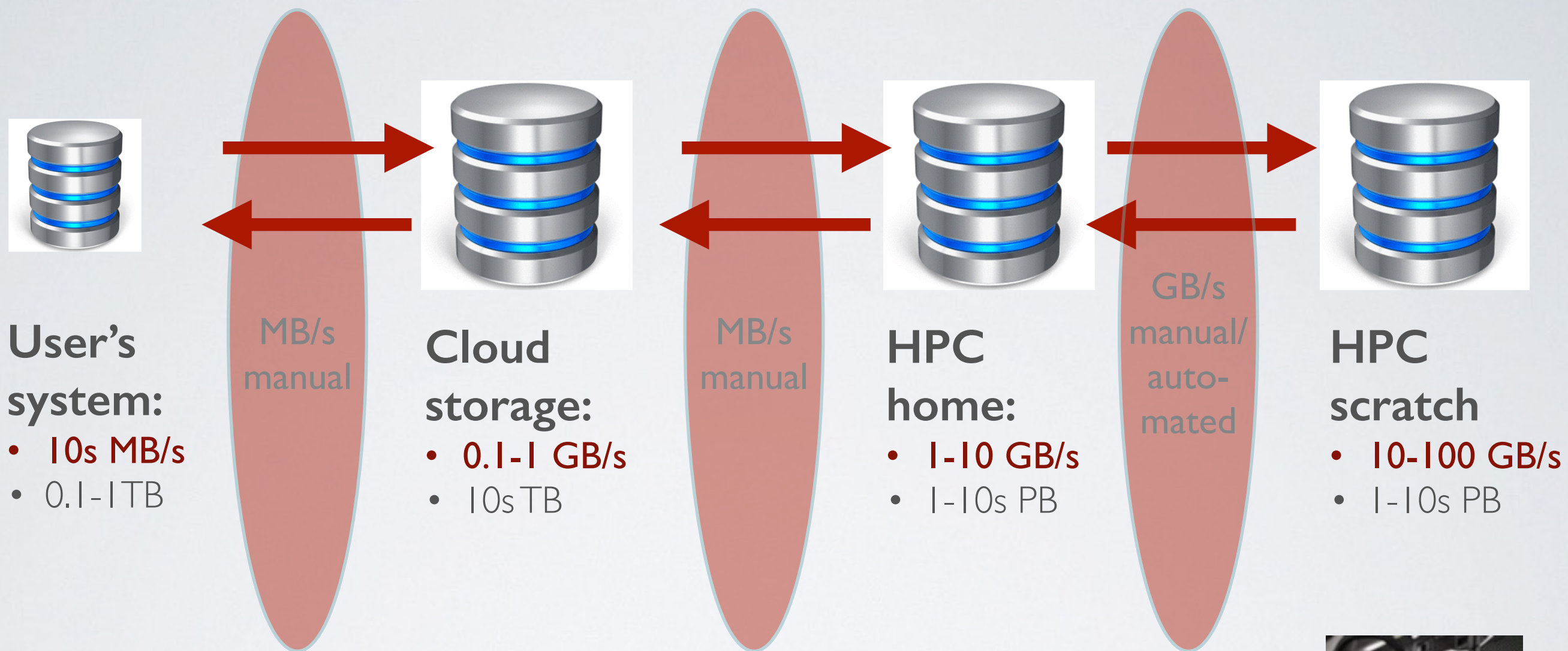
Storage



# TYPICAL HPC DATA FLOW:

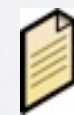
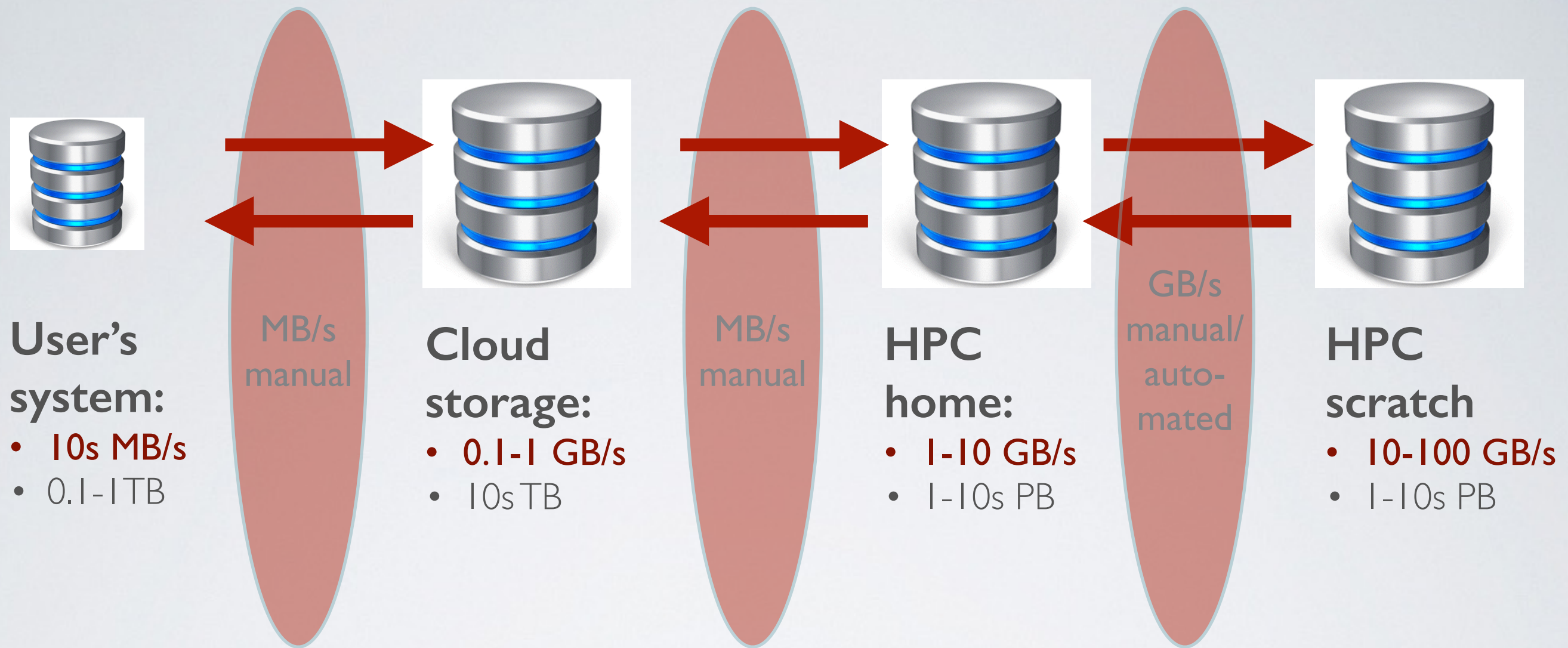


# HPC+CLOUD DATA FLOW:



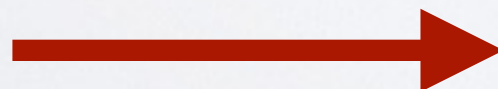


# OUR USE-CASE:



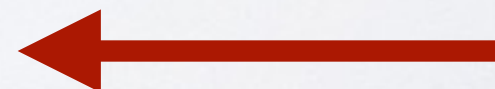
## Input data:

- ~3 GB
- 100 files



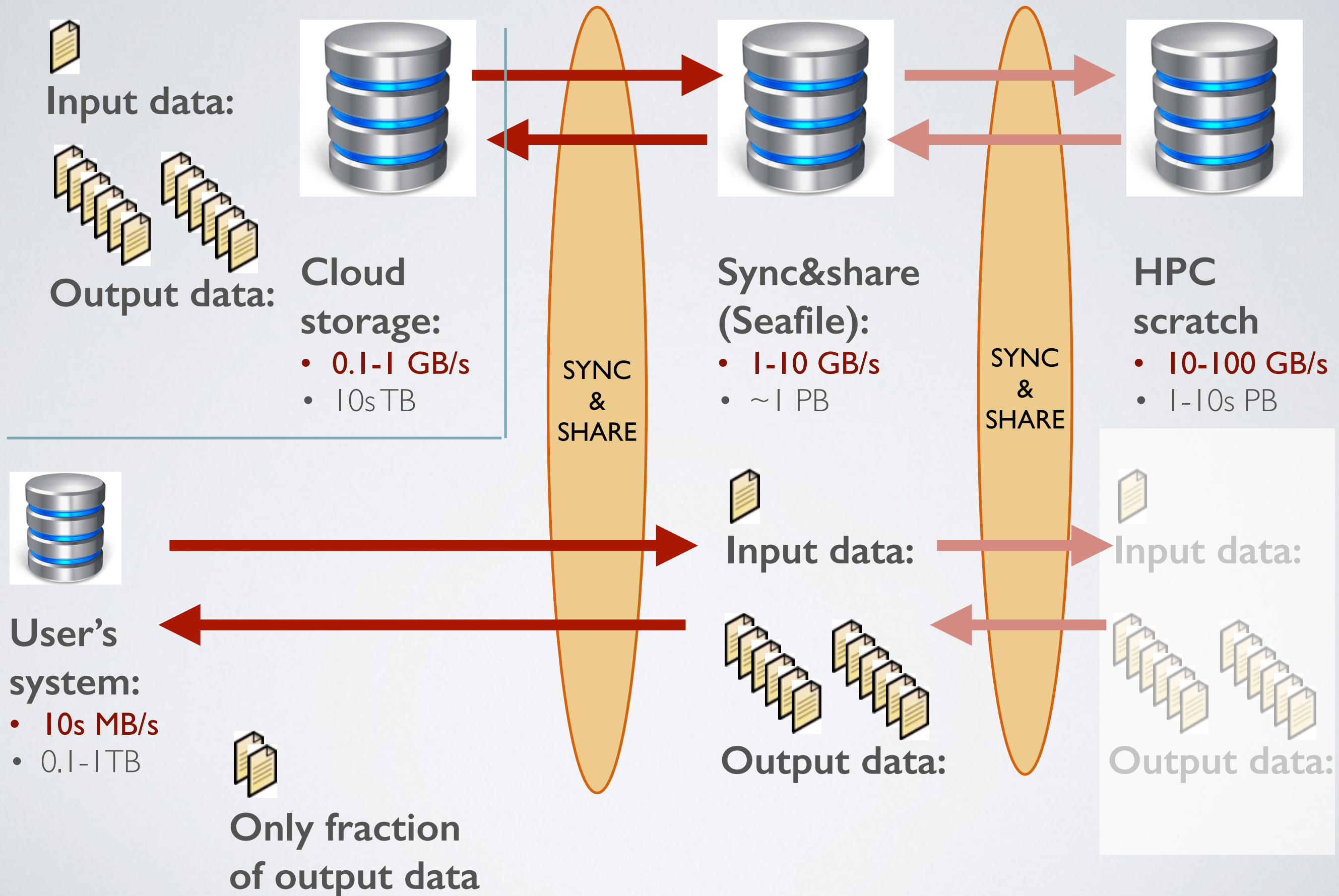
## Output data:

- ~40 GB per run & file
- 100 files





# POSSIBLE TARGET APPROACH





# TECHNOLOGY SELECTION

- **Requirements:**
  - **efficiency**
  - **reliability**
  - **robustness**
- **OK, but what system can cope with this?**

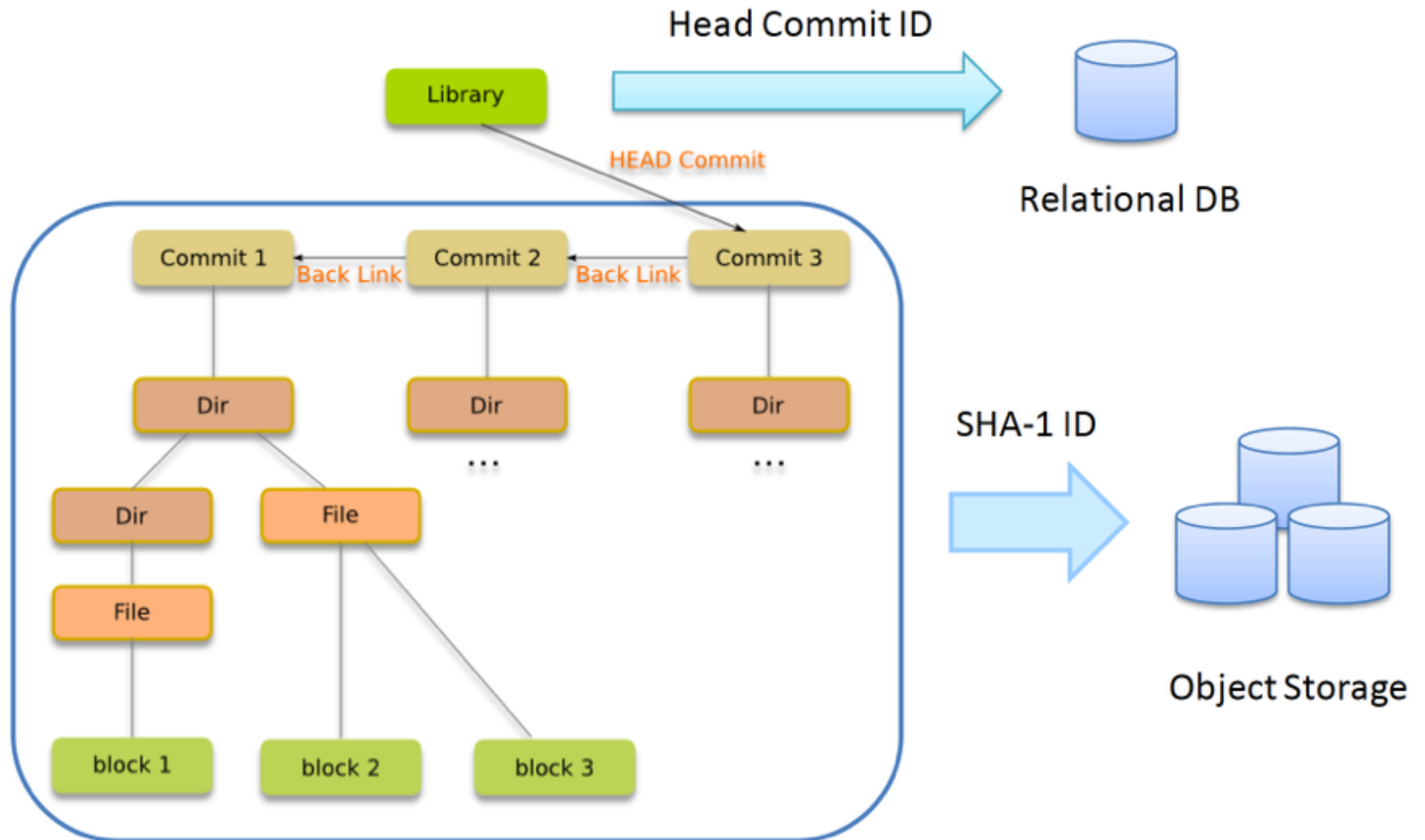




# WHAT IS SEAFILE?

- **Seafile is a sync & share system with focus on:**
  - **reliability** - data model, robust synchronisation algorithm
  - **efficiency** - low-level implementation (C), proper data model
- **Synchronisation mechanism:**
  - Snapshot based synchronisation (not per-file versioning)
  - Only deltas included in commits (content-defined chunking)

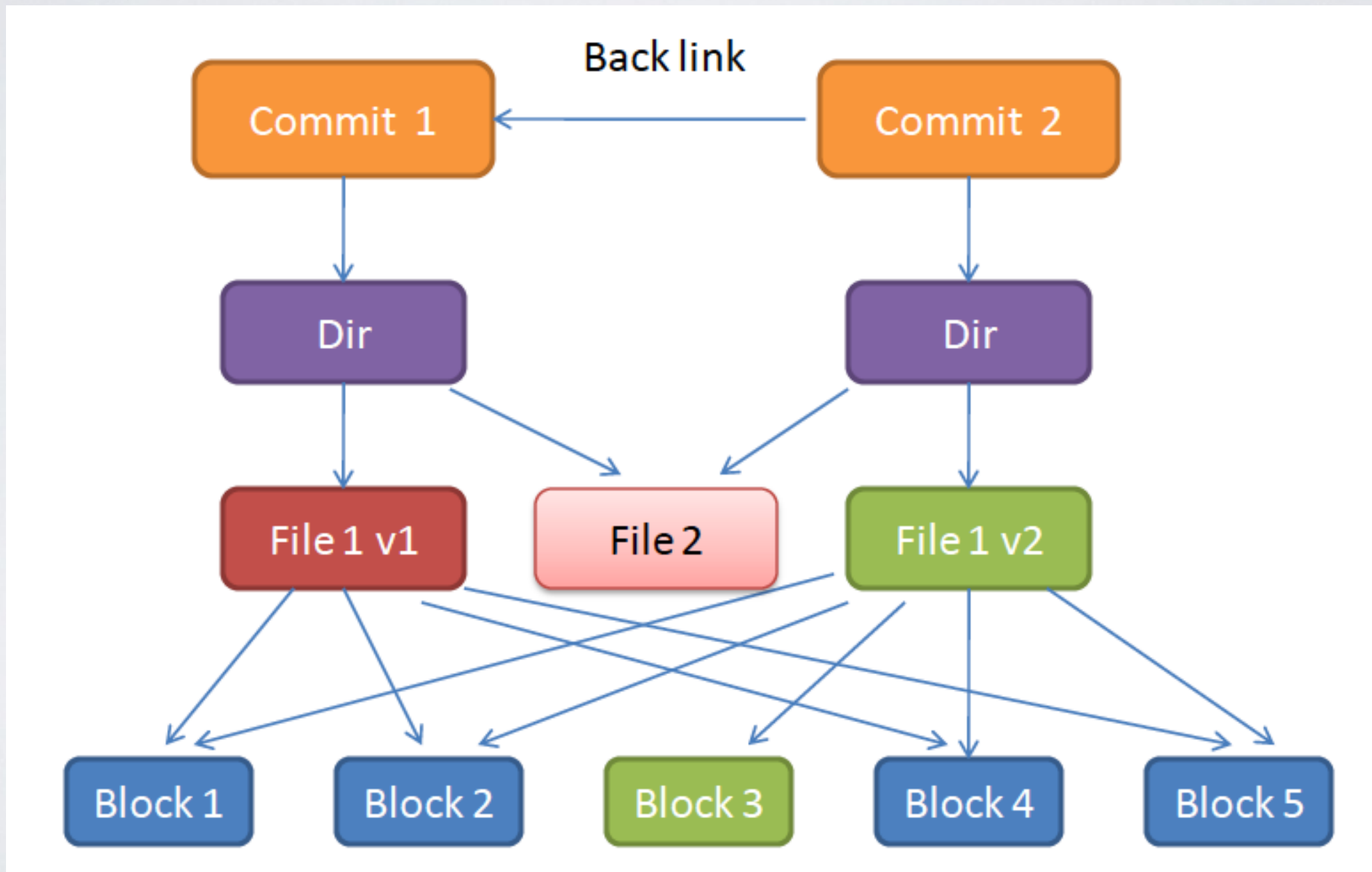
# SEAFILE SYNC MECHANISM: SNAPSHOT-BASED (NOT PER-FILE VERSIONING)





# SEAFILE SYNC MECHANISM:

ONLY DELTAS INCLUDED IN COMMITS,  
CONTENT DEFINED CHUNKING ALGORITHM USED FOR DEDUP





# FOCUS ON PURPOSE, NOT (TOO) MULTI-FUNCTION



Source: <http://www.fastcarinvasion.com/must-see-moment-tractor-crosses-way-racing-car/>





# SEAFILE'S STRENGTHS:

- **Well-optimized synchronization engine & architecture:**
  - **no overhead on CPU:**
    - well-optimised synchronisation engine
  - **no load on DBMS:**
    - minimum data in the DB (only shares, etc.),
    - metadata in storage backend
- **Seafile has potential to address I/O intensive workloads**
  - for large-scale sync & share services (this is what we do currently)
  - even as an HPC home (this is what we're testing)

# SEAFILE PERFORMANCE

2016 BENCHMARKS COMPARING SEAFILE WITH OTHER PRODUCTS

SPEED	Seafile [files-dirs/s]	theOther [files-dirs/s]	difference
Small files upload	627	27	23x
Small files download:	940	43	22x

SPEED	Seafile [GB/s]	theOther [GB/s]
Large files upload	0.17	0.11
Large files download	0.29	0.71

**SMALL FILES:** Linux kernel source v. 4.5.3: 706 MB of data, 52 881 files, 3 544 directories

**LARGE FILES:** 5 x 1GB

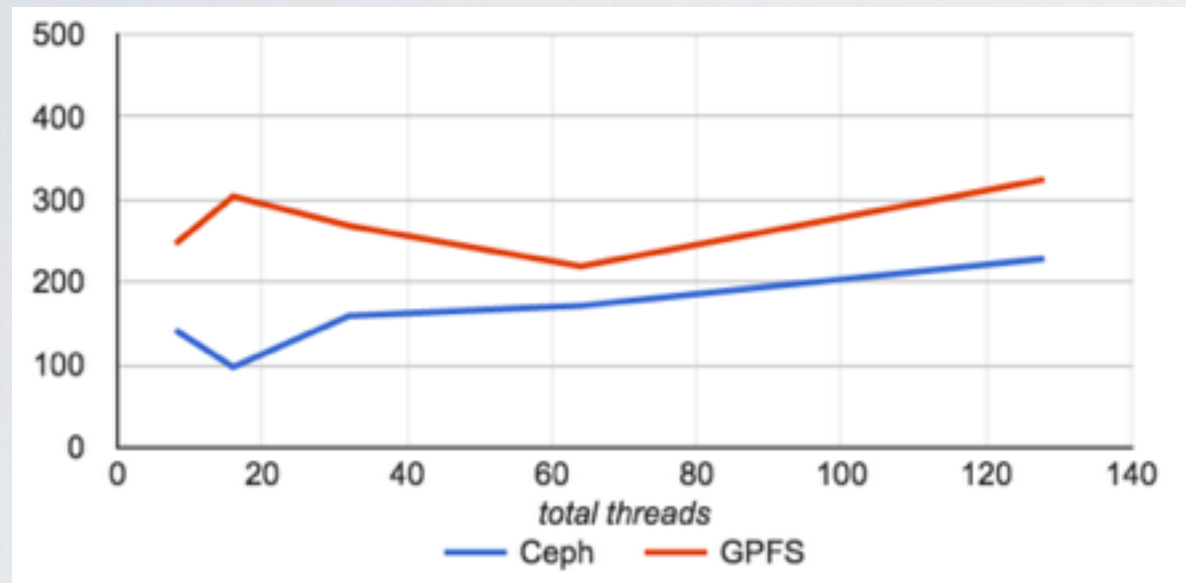


# SEAFIL PERFORMANACE

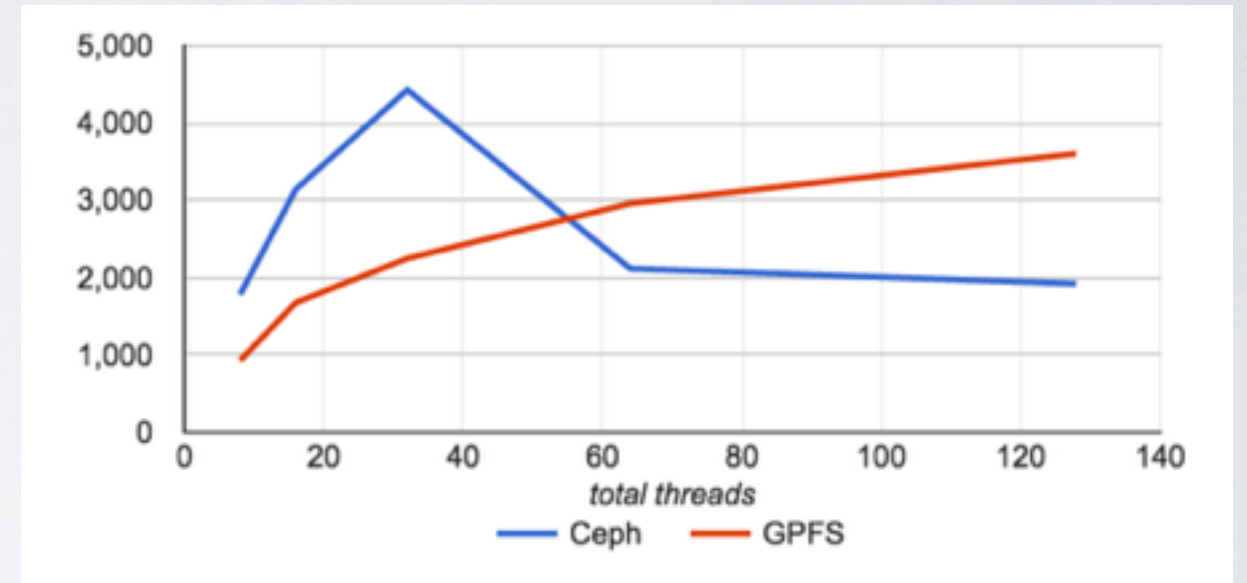
WITH CEPH AND GPFS BACKENDS (2017)

**SMALL FILES** TEST: 45K X 100KB FILES [FILES/S]

UPLOAD

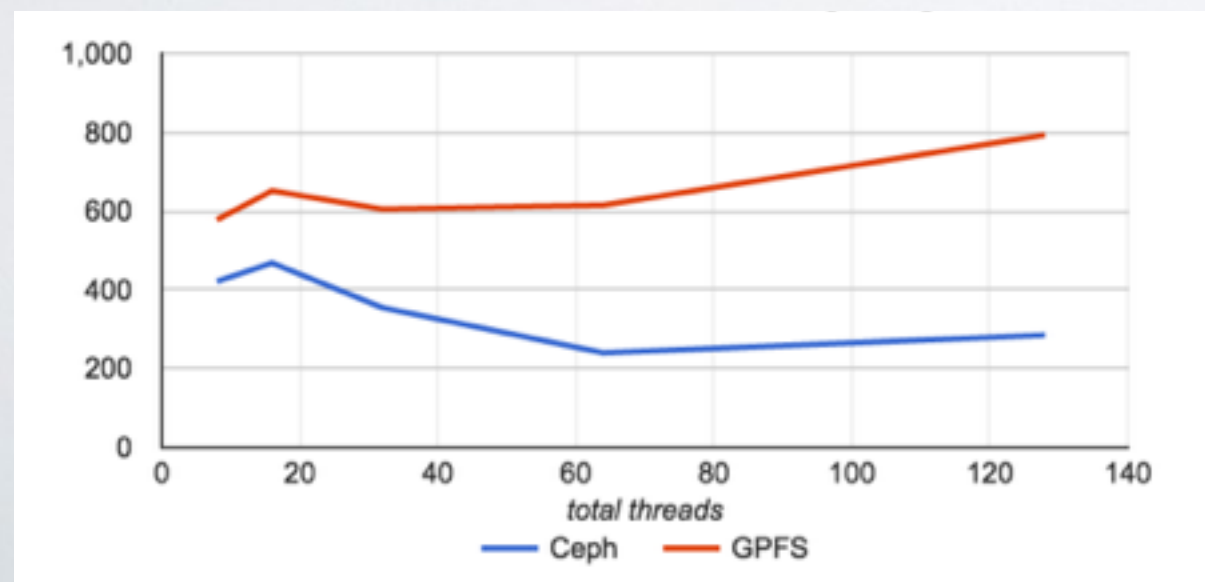


DOWNLOAD

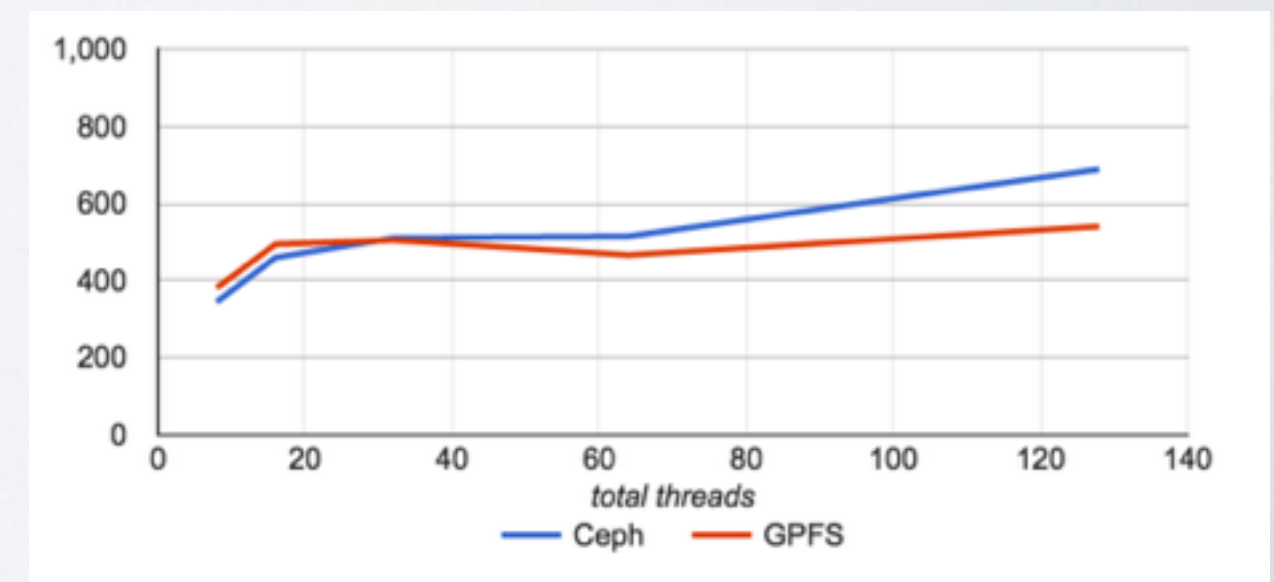


**LARGE FILES** TEST: 4,4 GB FILES: [MB/S]

UPLOAD



DOWNLOAD



# WHO WE ARE?



- Poznań Supercomputing and Networking Centre is one of the largest **HPC centres** in Poland
- It's also an **network** and **services provider (NREN)**; services include: cloud computing, storage, backup etc.



- **BOX** is a **country-wide sync & share service**:
- aimed at large user base (10s thousands); millions of files expected
- in production since 2015, PoCs with large universities ongoing



# HPC@



## ,EAGLE' CLUSTER:

- **1.4 PFlops** cluster
- 172th on Nov 2017 (Rmax: 1,372)
- **33k cores / E5-2697v3**
- **301 TB RAM**
- Infiniband FDR
- **Scratch** on **6PB**, 120GB/s **Lustre**
- **Homes** on **5PB**, 20GB/s **GPFS**







# - POLISH NREN & SERVICES PROVIDER

- **PIONIER NETWORK**

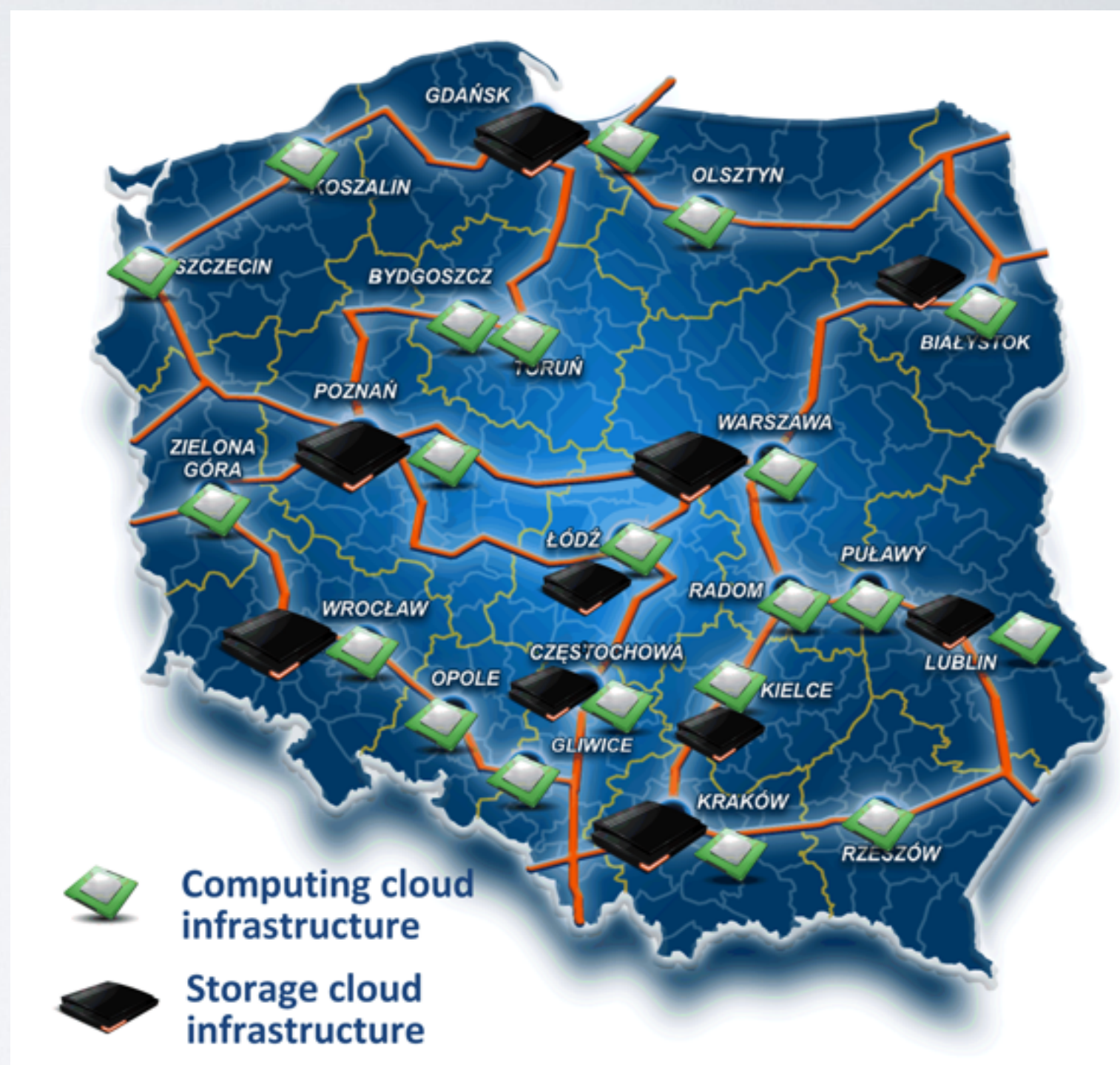
- **8000 kms of own fibers**
- 3500+ public institutions
- links to Geant, AMS-X, CERN

- **Archival Storage Services:**

- **14+PB** of space, 10 DCs
- 300+ client institutions
- Based on „National Data Storage” software developed in-house

- **Cloud computing services:**

- several 1000s of servers in 21 DCs
- 1000s of users







# IN DATA MGMT E-INFRASTRUCTURES

- **EUDAT:**  
Collaborative pan-European Data Infrastructure
  - PSNC delivers resources and services: B2SAFE, B2SHARE
  - R&D on object storage federations, HTTP-based federations
- **INDIGO-DataCloud:**  
European PaaS-based cloud for e-Science:
  - participating in work related to extending CDMI protocol/standard with QoS-related mechanisms
  - providing interfaces to object stores



- **GEANT**

- **Connectivity:**

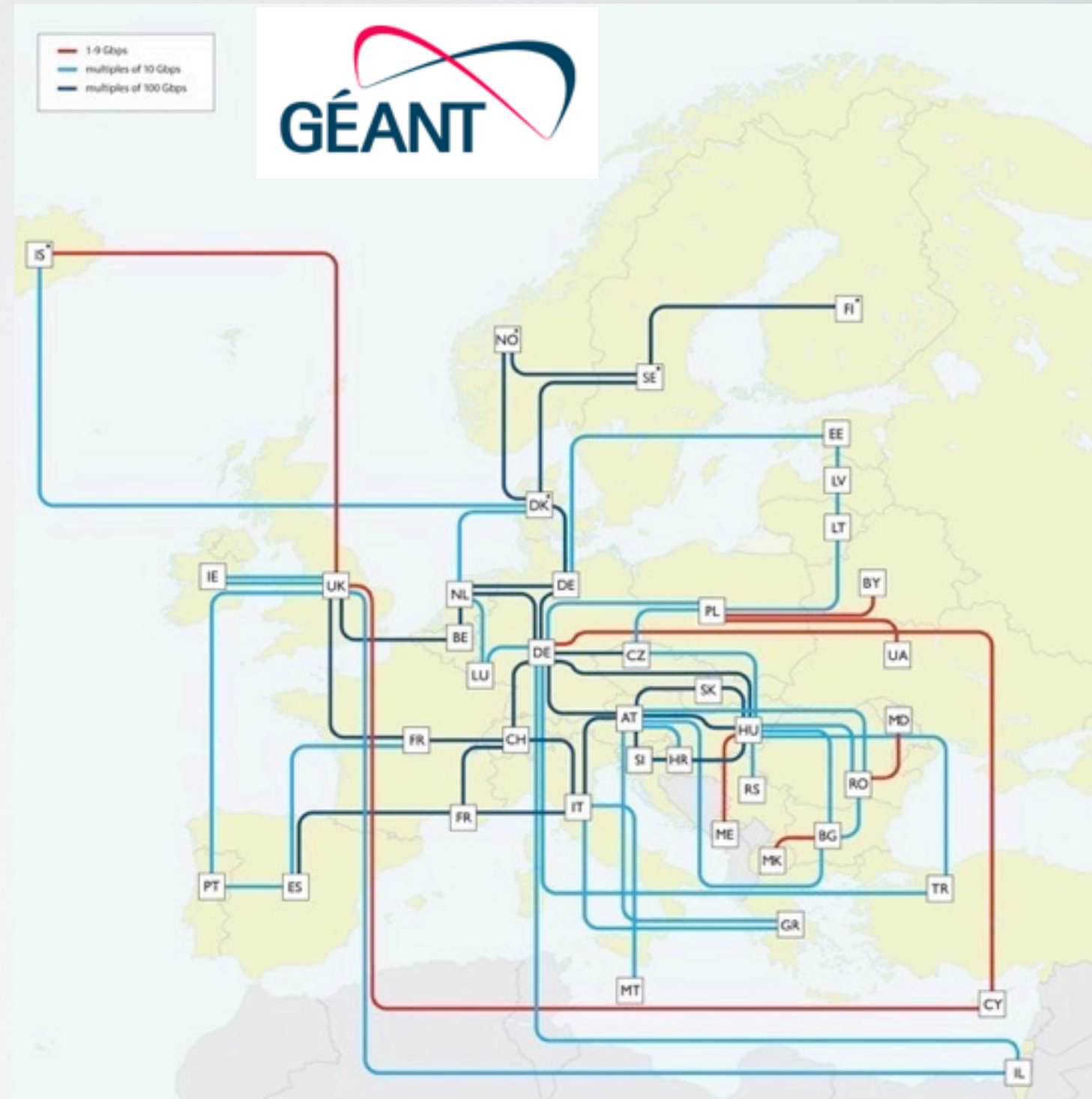
- multiple 10/100 Gbit lines

- **Collaborations: GN4 project:**

- software defined networks, infrastructure
    - multi-media, e-learning
    - cloud services incl. brokerage

- **Collaborations:**

- task forces: media, NOC etc.
    - special interest groups: cloud services & software stacks







[BOX.pionier.net.pl](http://BOX.pionier.net.pl)

# BOX@PIONIER



Country-wide sync&share service

- aimed at large user base (100s thousands);
- millions of files expected
- in production since 2015
- PoCs with large universities ongoing

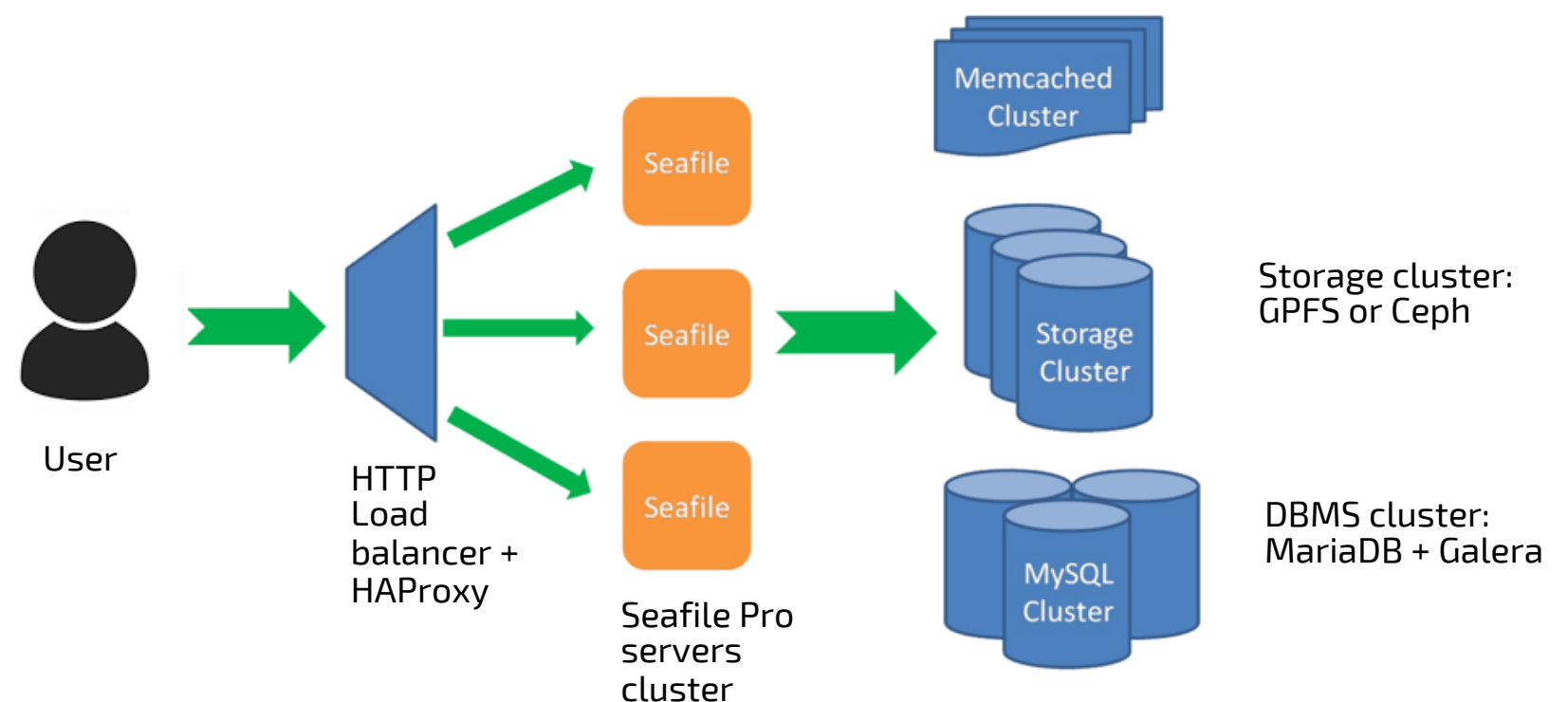


[BOX.pionier.net.pl](http://BOX.pionier.net.pl)

# TEST INFRASTRUCTURE

We used production **BOX@PIONIER** service:

- Based on cluster of Seafile Pro servers
- Database cluster: MariaDB + Galera
- HTTP load-balancers HTTP: HAProxy etc.
- Storage cluster: GPFS (GPFS+cNFS)



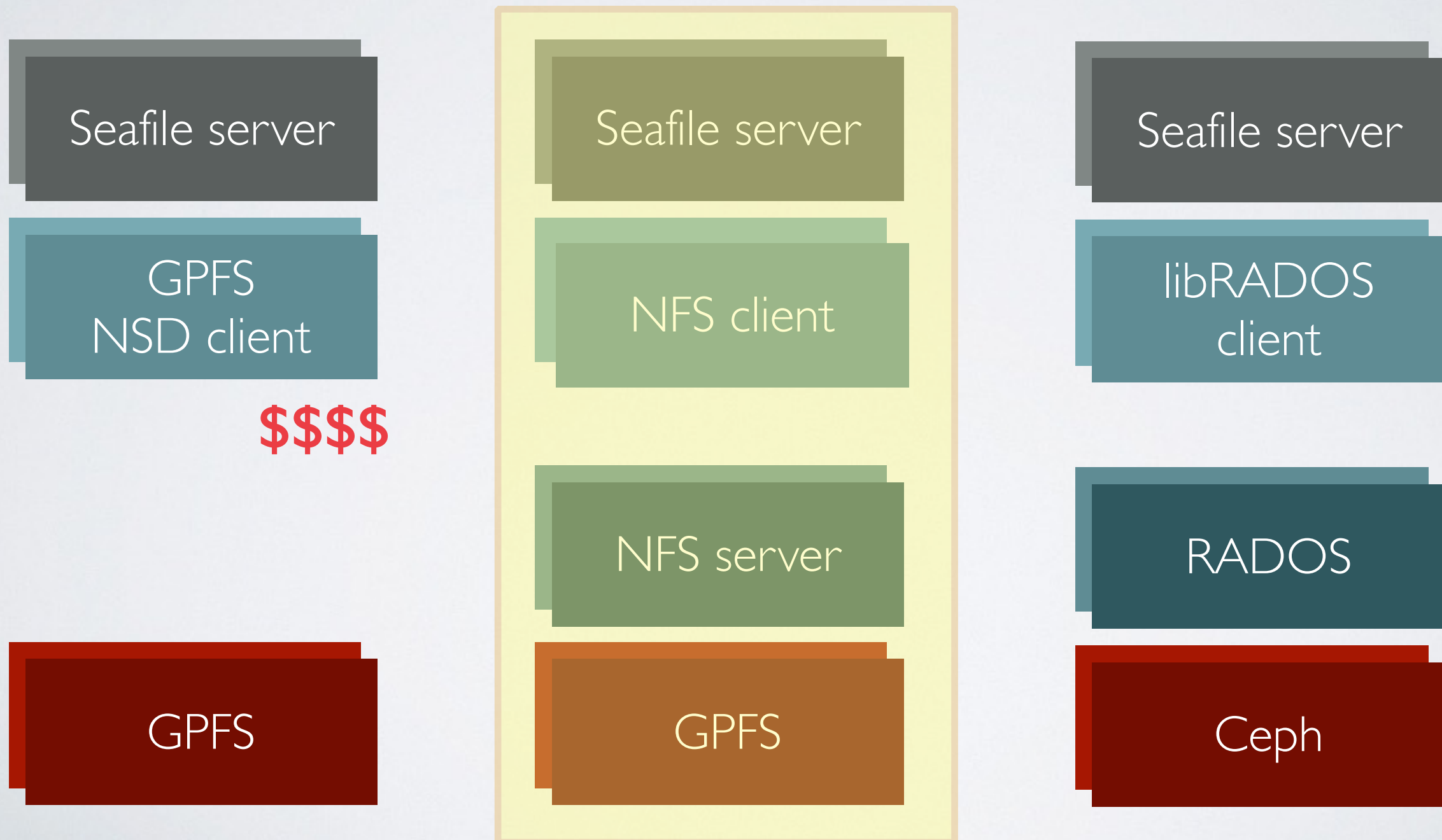




# TEST INFRASTRUCTURE

We use GPFS at the Seafire back-end through the cluster NFS gateways

We tested Ceph but GPFS performed better



# DETAILED CONFIGURATION



## Seafile:

- 1x load-balancer
- 2 Seafile servers
- Maria DB Galera
- MemcacheD
- Storage back-ends:
  - Ceph
  - GPFS

## GPFS back-end:

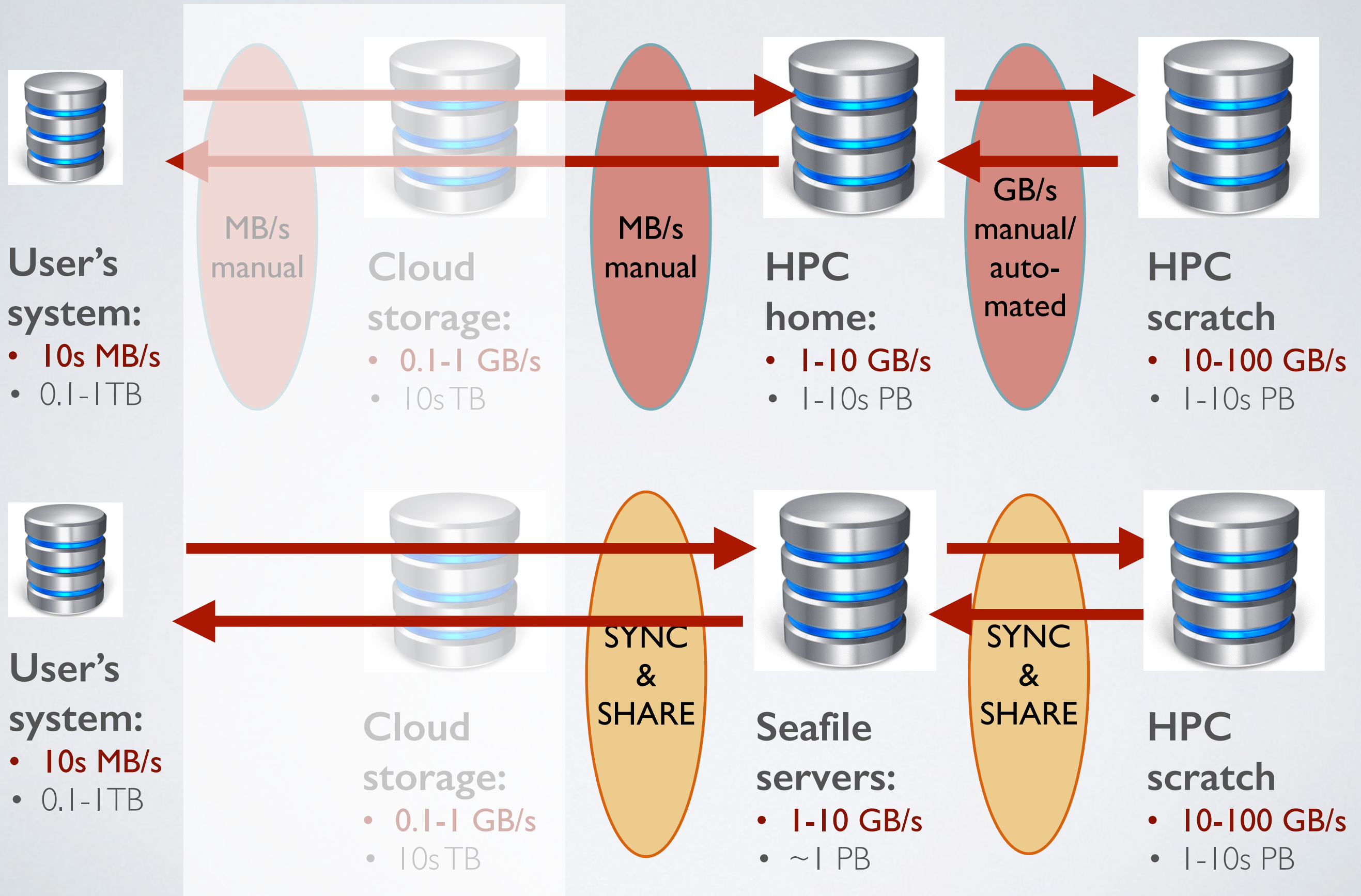
- **2 GPFS servers**
  - 2 CPUs, 128GB RAM
  - 2x 10GbE, 2x 10GbFC
  - GPFS v4.2.2
- **Disk array:**
  - Huawei OS 5500
- **120 HDDs, RAID6**
- Interface:
  - 2x10GbE
  - NFSv3

## HPC storage:

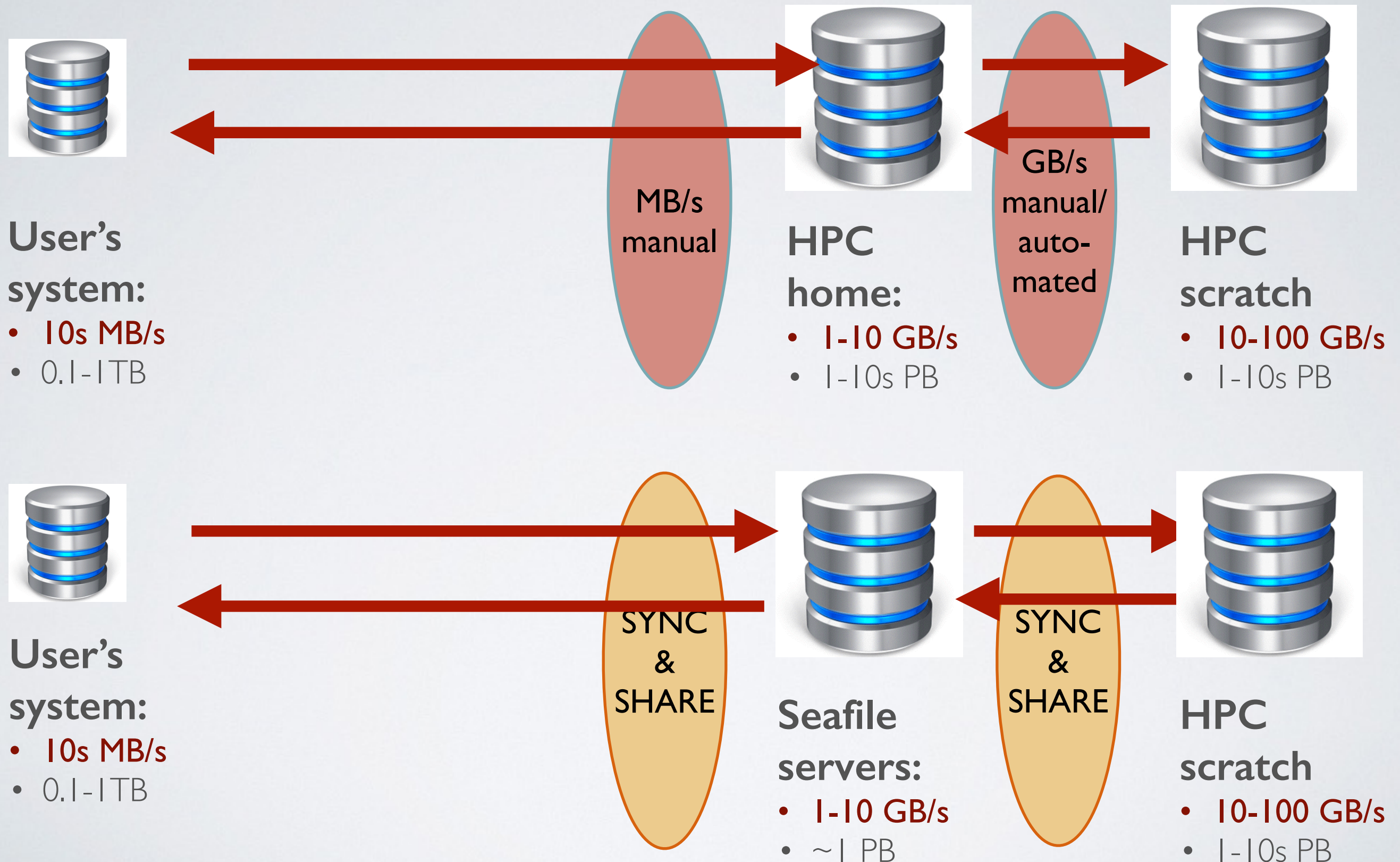
- **Scratch / Lustre:**
  - 16 OSS servers
  - 2 MDS servers
  - 5,6 PB physical space
- **Home/GPFS:**
  - 8 NSD servers
  - 8 cNFS servers
  - 2 PB physical space
- **Interfaces**
  - 1 Gbit link to „world” :(



# TEST PROCEDURE & RESULTS

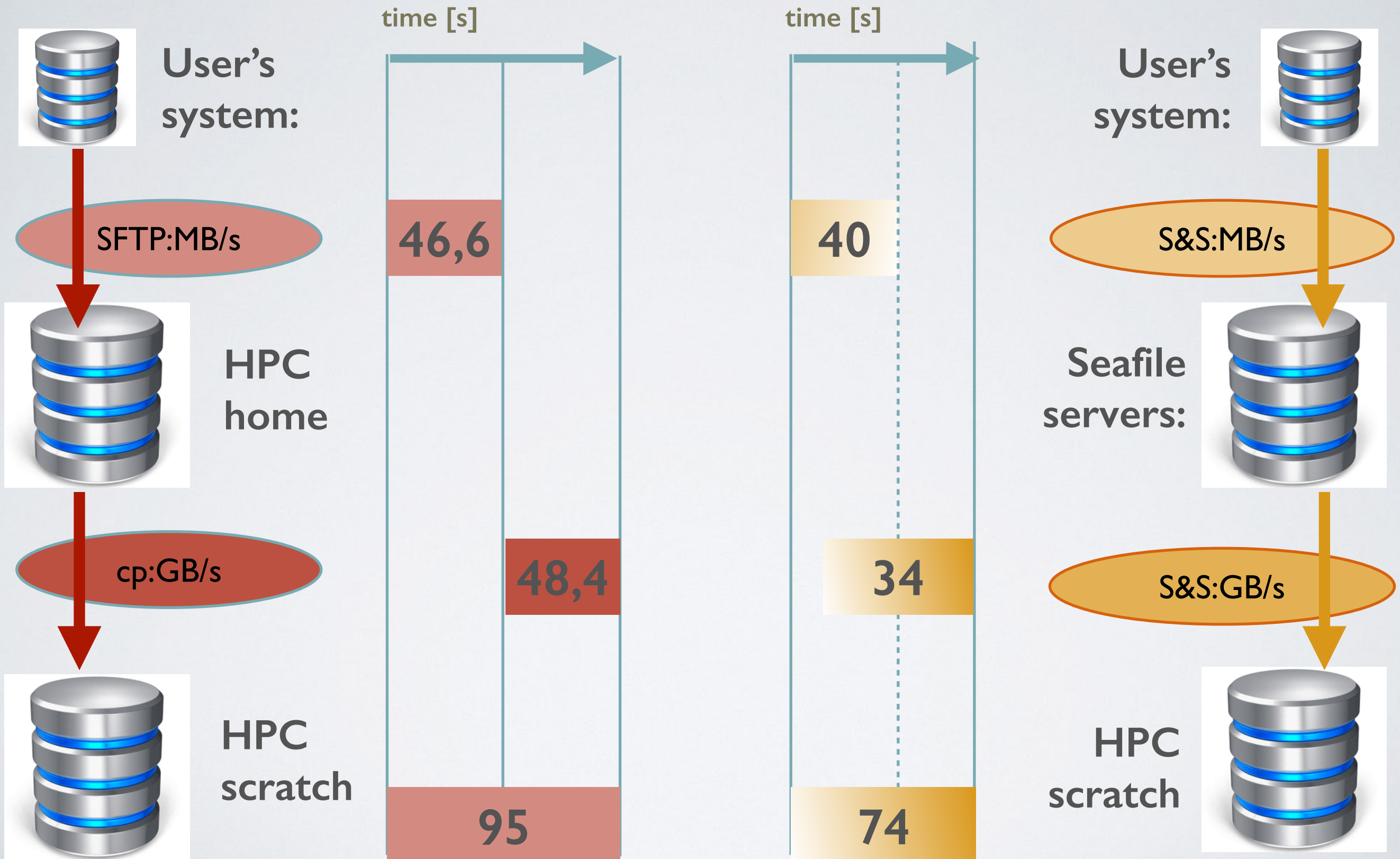


# TEST PROCEDURE & RESULTS

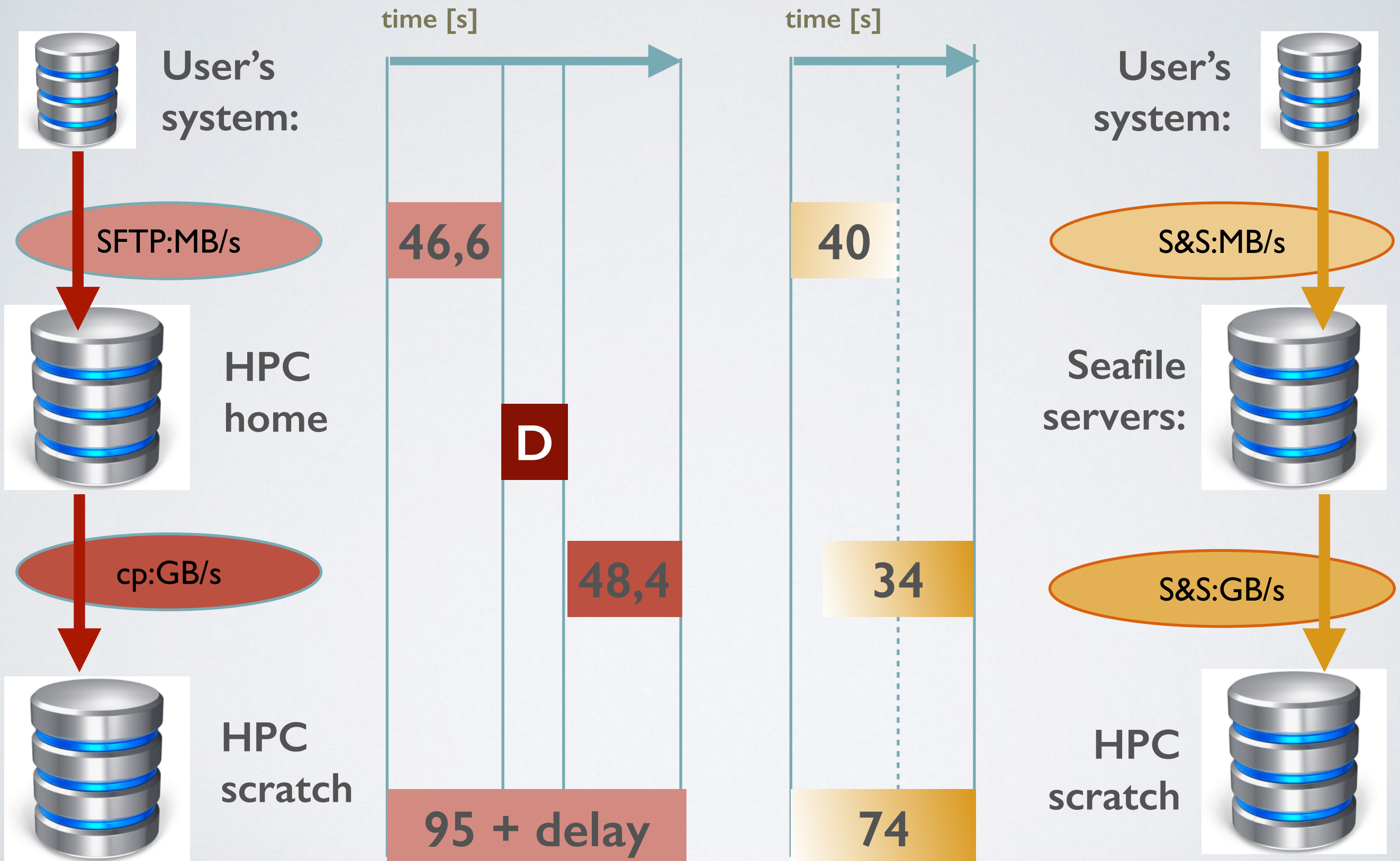




# TEST PROCEDURE & RESULTS



# TEST PROCEDURE & RESULTS







# OBSERVATIONS

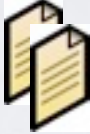


- **The approach** proposed is **convenient for users and efficient**:
  - ease of use + **automation** of data movement
  - **no delays** on user's / workflow engine reaction
  - ***instant availability** of results as they are produced*
- **Admin's point of view**:
  - Space usage is similar to storing data in ,classical' home
  - Only minimum extra work required
  - Users are happy so admins are
- Overall this seems to be a way to go in some use-cases

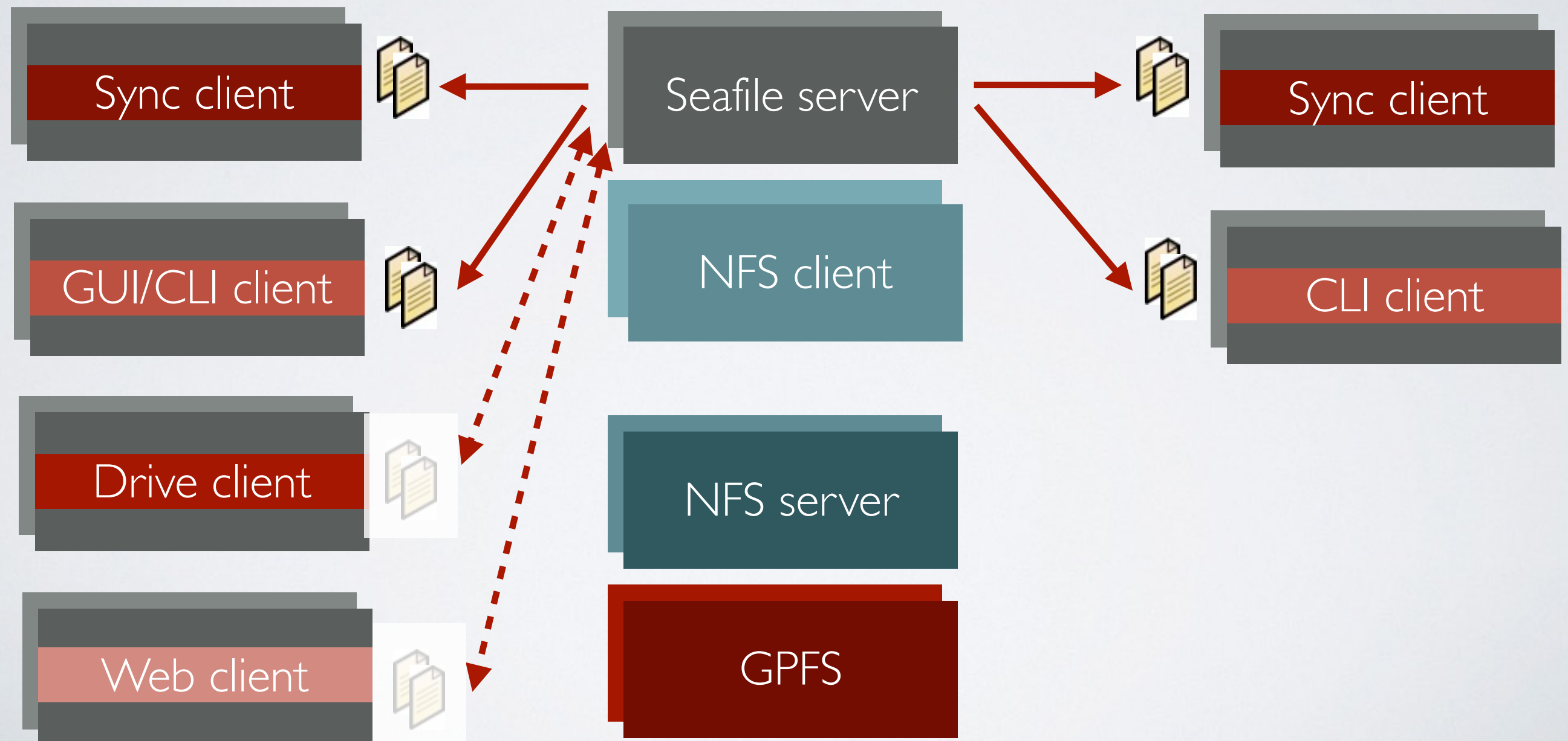


# STORAGE USAGE

User's system  
or VM in cloud

Sync & share  
system 

HPC system







# LIMITS/FUTURE (I)



- **Limitations** of our approach:
  - Only 1 Gbit link to cluster storage - we used temporary/testing setup
  - More interfaces/ machines needed - ,gateways' in a systematic approach
- **Only synchronisation client tested:**
  - Virtual drive-like access possible: **Seadrive** could be used for **ad-hoc access** without copying them to local drives within the user workstation and/or in VM embedded in cloud



## LIMITS/FUTURE (2)



- User credentials ,delegation':
  - for testing we created a dedicated ,a group' account in Seafile; users joined to be able to share data while keeping their private credentials for themselves
  - more systematic approach to configuring user accounts would be required in future (ideally ,common' accounts in sync & share and in the HPC cluster)
- Configuration of directories/libraries to be sync'ed:
  - performed manually for testing purposes
  - more automated approach needed





# SUMMARY



- We tested in practice a concept of replacing ,classical' home filesystem with sync & share solution
  - this approach is a work in progress, with limitations
  - however looks like the way to go in some use-cases
  - proper sync engine provides performance relevant to HPC
  - for now: ,zero development' approach
  - more systematic approach may require development work



SYNC&SHARE TO  
REPLACE HPC HOMES?

THANK YOU!  
QUESTIONS?

Thank you!

Credits to:

Krzysztof Wadówka,  
Maciej Brzeźniak,  
Marcin Pospieszny,  
Piotr Brona,  
Radosław Januszewski  
HPC Department

