SYNC&SHARE SOLUTION FOR MASSIVE MULTIMEDIA DATA

Maciej Brzeźniak, Krzysztof Wadówka, Paweł Woszuk
HPC Department

Maciej Głowiak, Maciej Stróżyk
New Media Department

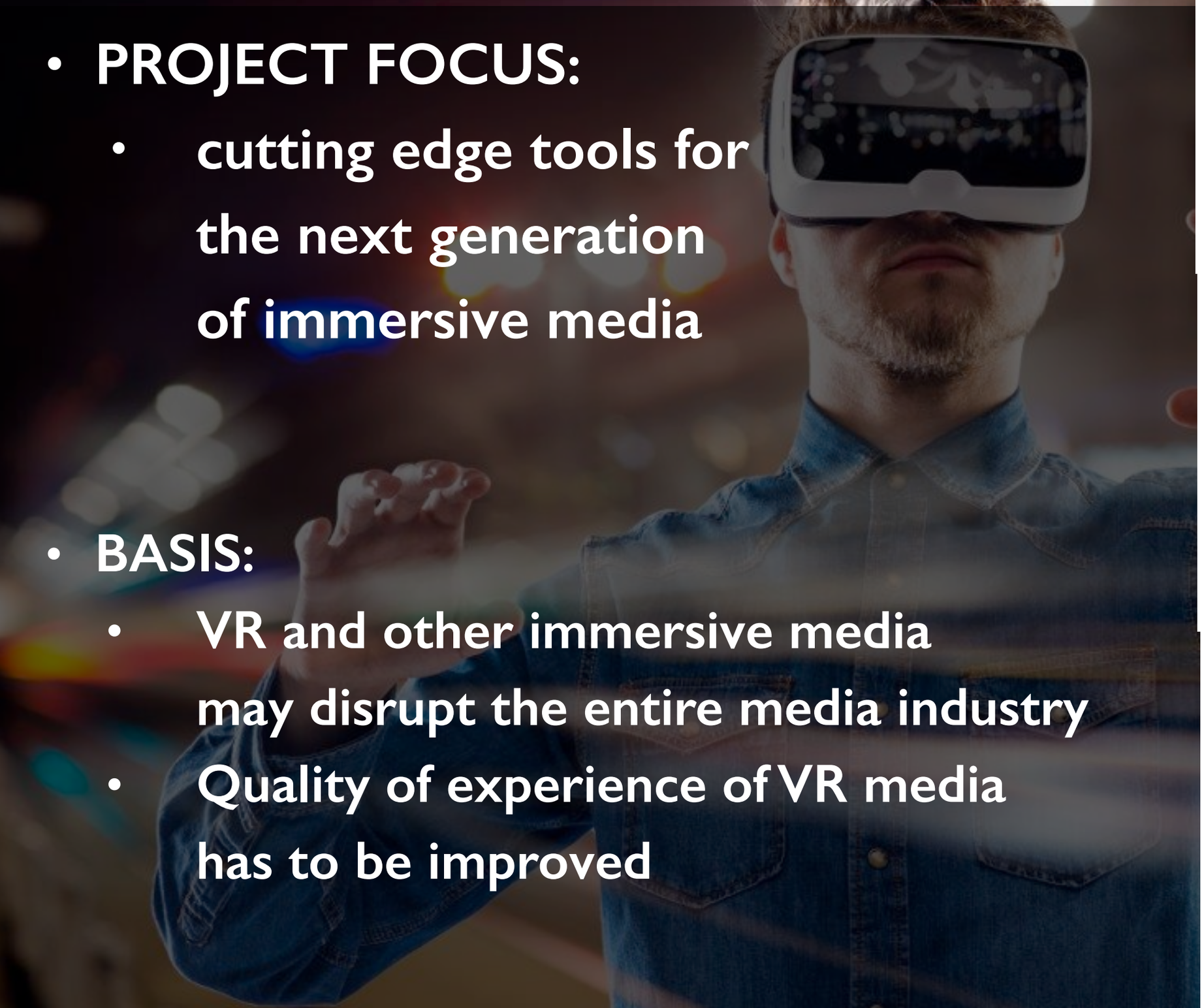
PSNC

AGENDA

- Context + massive multimedia data challenge
- Solution
- PSNC & BOX: who we are, why we are doing this
- Future work (work in progress)
- Observations

Context

CONTEXT: IMMERSIFY



- PROJECT FOCUS:
  - cutting edge tools for the next generation of immersive media

- BASIS:
  - VR and other immersive media may disrupt the entire media industry
  - Quality of experience of VR media has to be improved

PSNC

spin digital

ARS ELECTRONICA FUTURELAB

MARCHÉ DU FILM
FESTIVAL DE CANNES

NORRKÖPING VISUALIZATION CENTER

CONTEXT: IMMERSIFY

- **DETAILED GOALS:**
  - (1) *develop* advanced **video compression technology** tailored for the needs of the VR video

    enabling delivering and display the huge files that will appear as a result of increased *resolution*, *frame rate* and better *image formats*.
  - (2) *allow* the **widespread of immersive content**, **and facilitate its distribution and exhibition**

    by supporting multiple devices and environments such as PC- and mobile-based head mounted displays, multi-display systems, and dome, immersive cinemas and deep spaces.
  - (3) a*llow* content creators to **produce highly personalized** content

    with seamless interactivity by developing the required tools to combine high quality video, 2D/3D CGI, and interactive elements.

MASSIVE MULTIMEDIA (MM) DATA

- **OVERALL DATA MGMT CHALLENGE:**
  - growing **volume**: **PetaBytes**
  - pressure for **performance**: GB/s, IOPS
  - user expectation for **ease of use**

- **MM DATA MGMT CHALLENGE:** 4k VIDEO **uncompressed**:
  - volume:
    - ~200MB / frame, 60fps:
    - **11 GB**/second - **703 GB**/minute - **41,2 TB**/hour
  - data flow:
    - content produced at PSNC (Poznań)
    - codecs developed and tested at Spin Digital (Berlin)

EXPECTATIONS:

- **SEAMLESS AND EASY DATA EXCHANGE**
  - **multiple iterations of the worklow**
  - **ad hoc data access -> filesystem like access**
  - *the less manual work the better*
- **ROBUSTNESS:**
  - with so many files (>200k / hour)
    we can't tolerate failures in copying
- **PERFORMANCE:**
  - should enable running tests of codecs
    without waiting the hours for access

The solution

WHAT IS SEAFILE?

- **Specialised** solution **designed for sync & share**
  - **reliable** - data model, synchronisation algorithm
  - **effective** - low-level implementation (C), proper data model
- **Backends** supported:
  - Filesystem, **NFS**, etc.
  - S3, Swift / **Ceph**

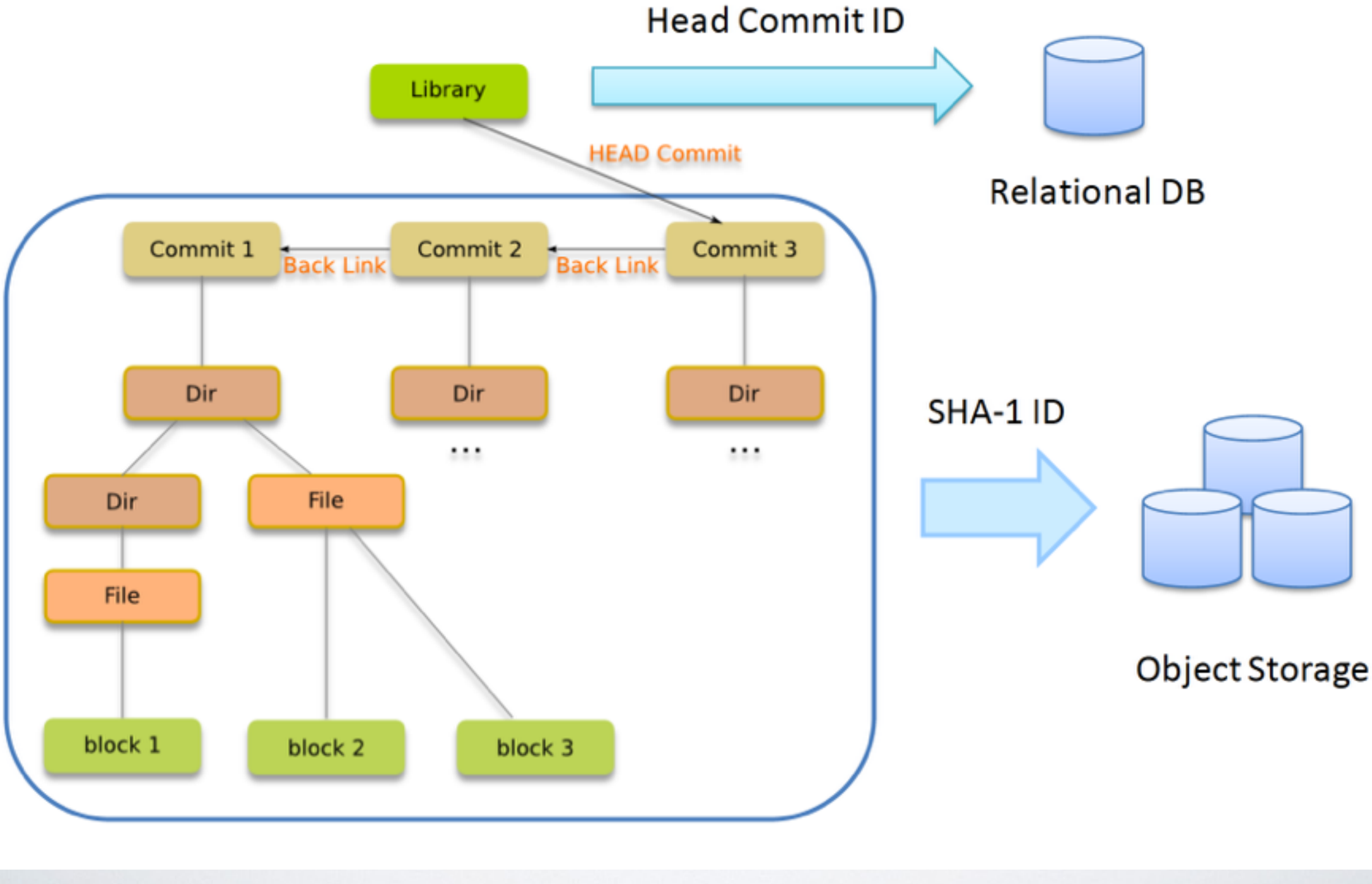
Seafile™

FOCUS ON PERFORMANCE,
AND RELIABILITY

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

SEAFILE SYNC MECHANISM:

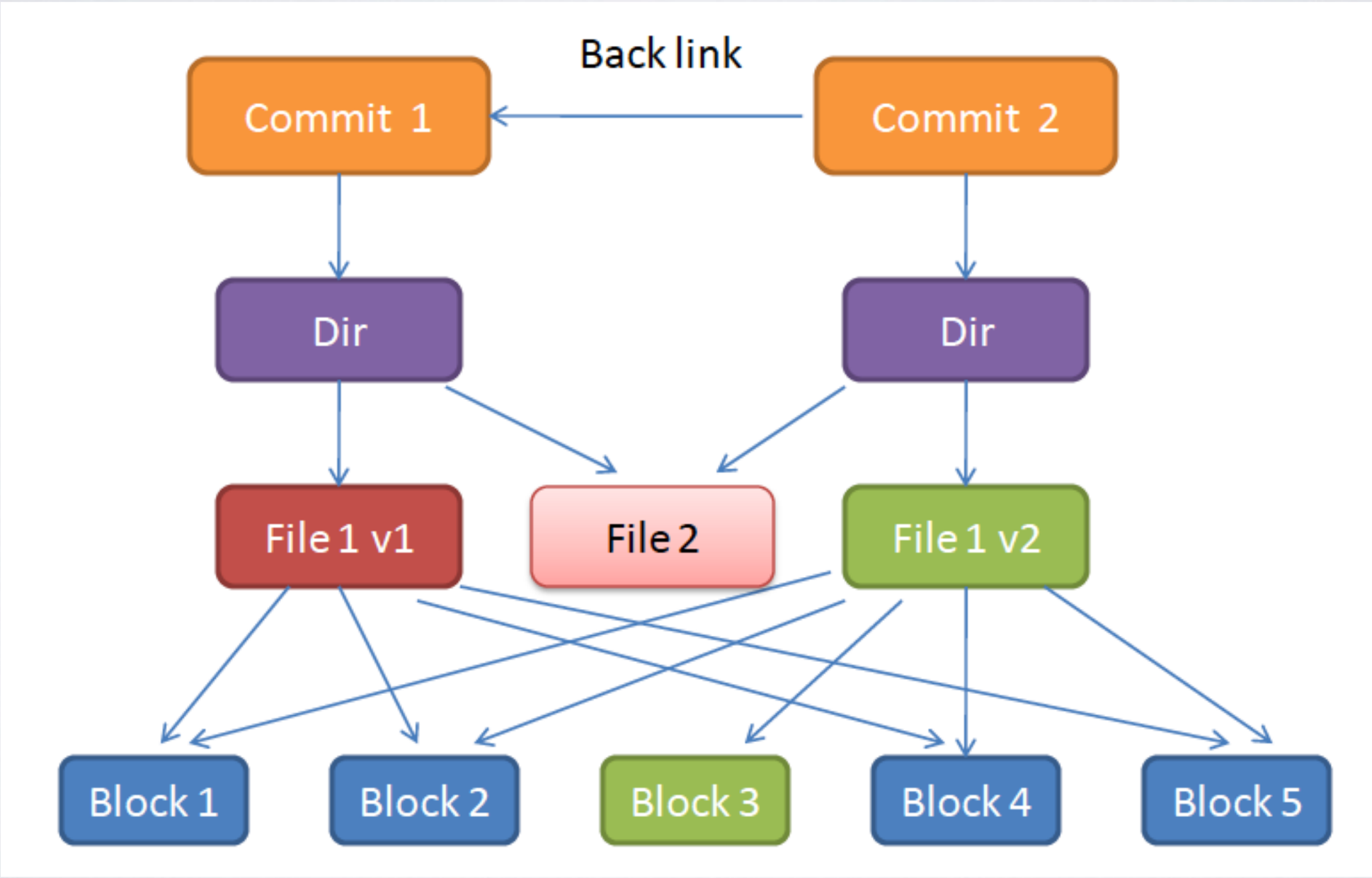
SNAPSHOT-BASED (NOT PER-FILE VERSIONING)



SEAFILE SYNC MECHANISM:

ONLY DELTAS INCLUDED IN COMMITS,

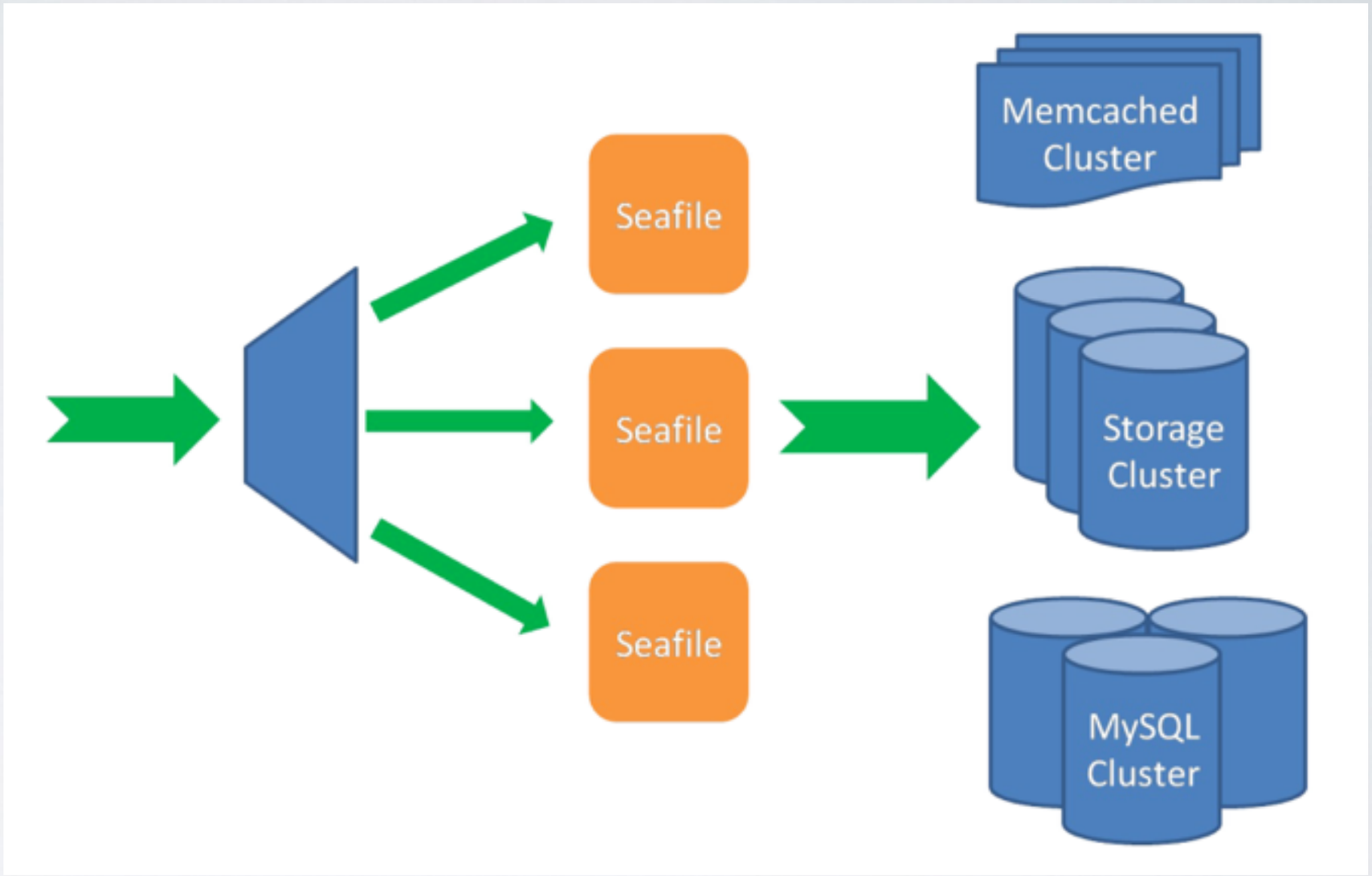
CONTENT DEFINED CHUNKING ALGORITHM USED FOR DEDUP



LOAD-BALANCED SETUP

Architecture:

- Load-balancer
- Seafile servers
- Storage back-ends:
- Memcached
- MySQL/Maria DB



Architecture scales horizontally

- Seafile application servers work independently
- They share minimum information through memcached

SEAFILE PERFORMANCE

LARGE FILES*) PERFORMANCE TEST (2016)

SPEED	Seafile [GB/s]	theOther [GB/s]
5x1GB file upload	0.17	0.11
5x1GB file download	0.29	0.71

LARGE FILES *)

- 5 GB file

SEAFILE PERFORMANCE

SMALL FILES*) PERFORMANCE TEST (2016)

SPEED	Seafile [files-dirs/s]	theOther [files-dirs/s]	difference
Client 1: upload	627	27	23x
Client 2: download:	940	43	22x

SMALL FILES *)

- Linux kernel source v. 4.5.3
  - **706 MB** of data
  - **52 881** files
  - **3 544** directories

SEAFILE 5. COMMUNITY, SINGLE 2-CPU SERVER, 120-DISK FC ARRAY, EXT4

SEAFILE VS OTHERS

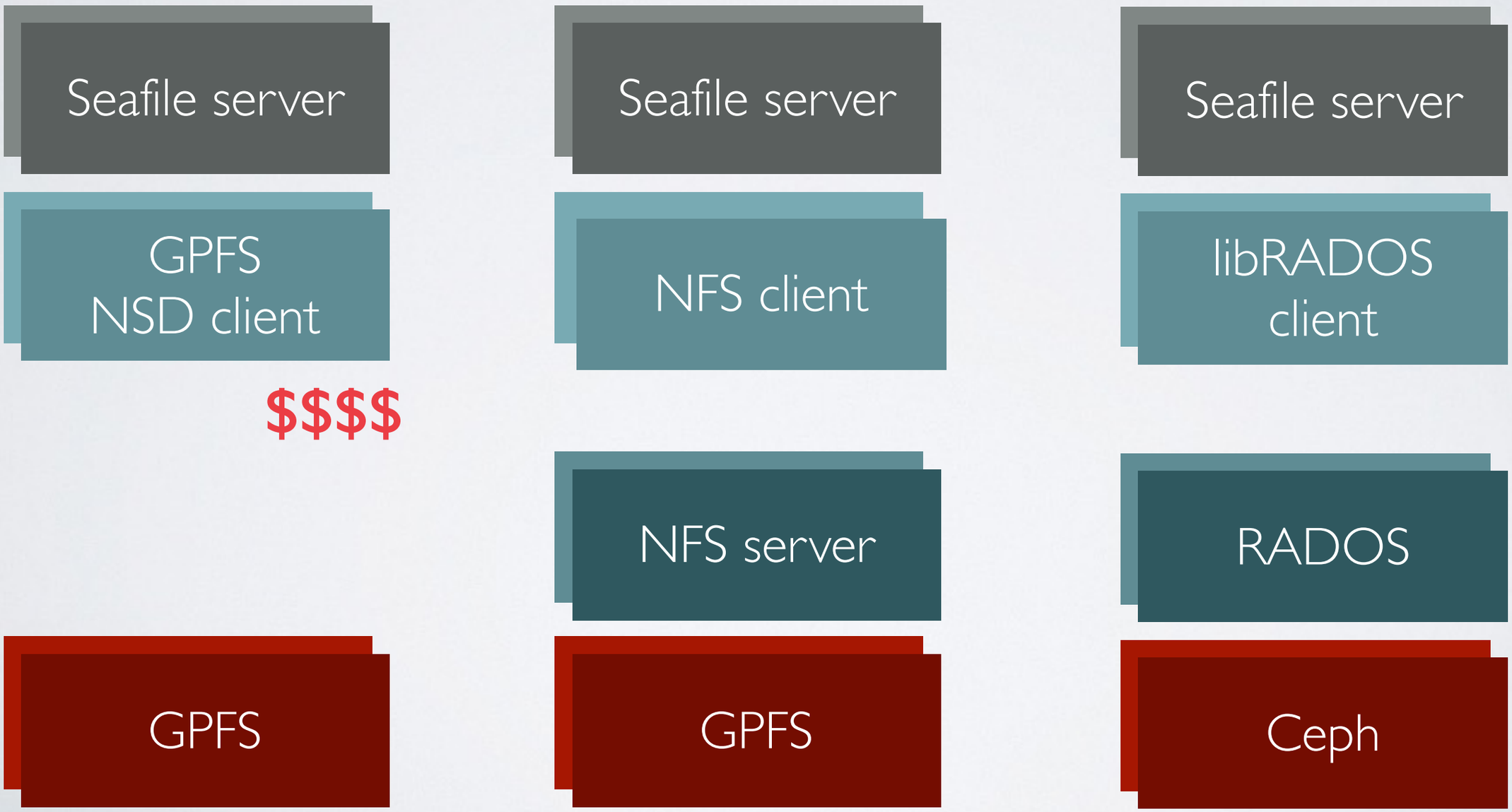
SMALL FILES PERFORMANCE TEST (TIME)

test	2016 test single Seafile server, very small files - Linux kernel source		2017, clustered Seafile 100kB files
SPEED	Seafile [files-dirs/s]	theOther [files-dirs/s]	clustered Seafile [files-dirs/s]
Client 1: upload	627	27	400
Client 2: download:	940	43	3400

BACKENDS FOR box

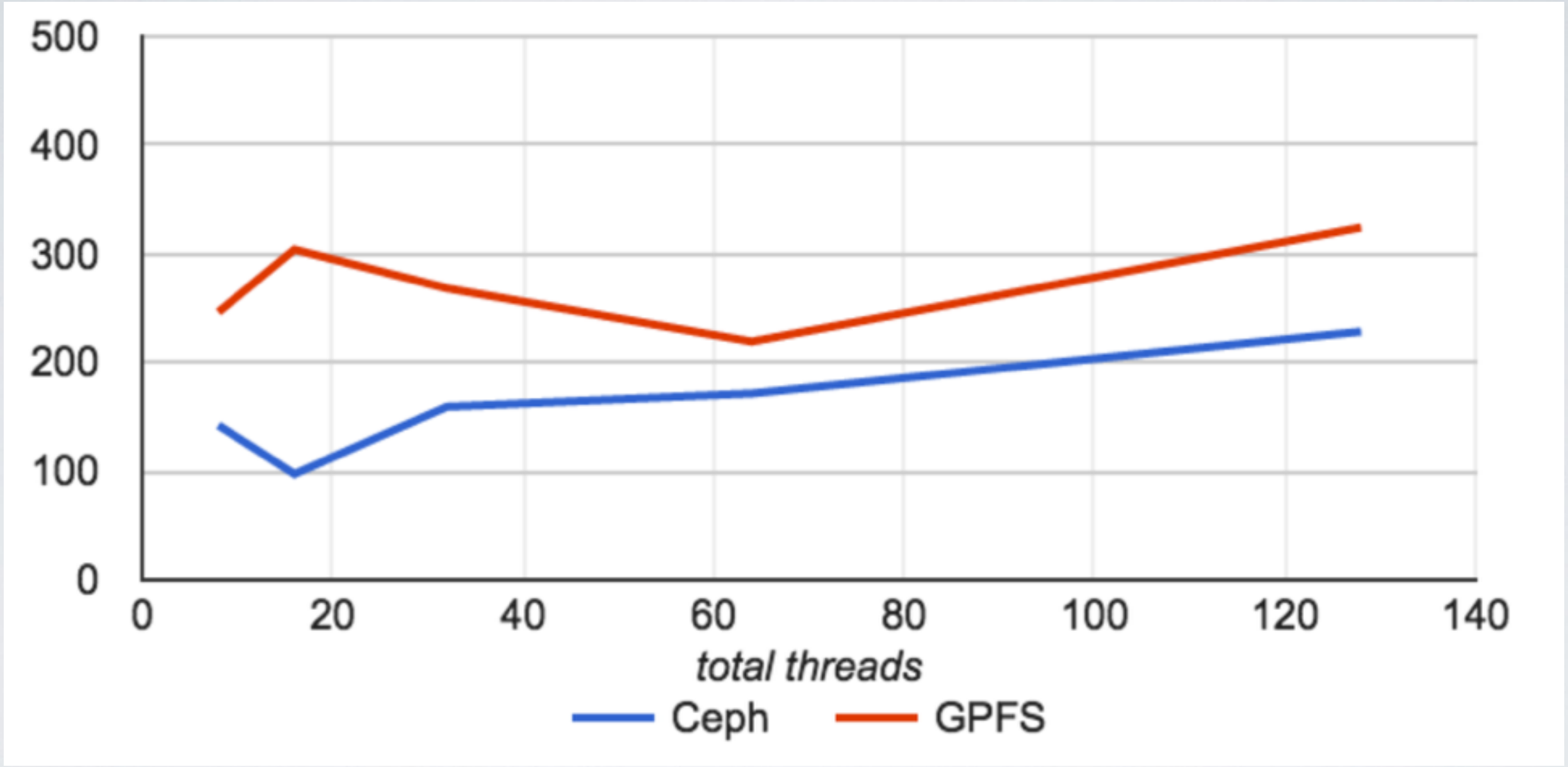
Having paid IBM already for GPFS
use them for sync & share?

Use Ceph
as everybody does ;) ?



UPLOAD RESULTS [FILES/S]

SMALL FILES TEST (45K X 100KB FILES)

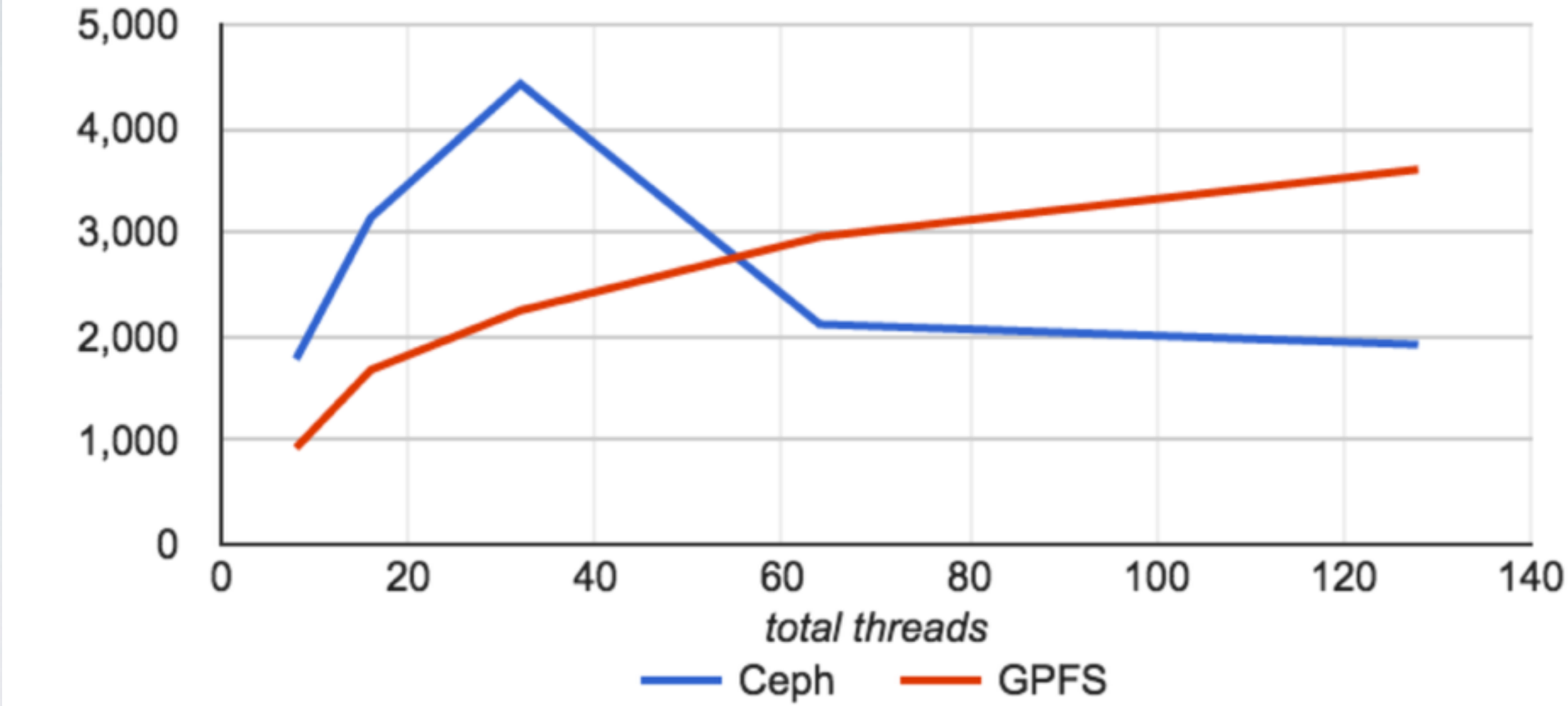


GPFS is up to 1.5-3x faster than Ceph:

3x replication in Ceph + intermediate storage step at Seafile server's back-end

DOWNLOAD RESULTS [FILES/S]

SMALL FILES TEST (45K X 100KB FILES)



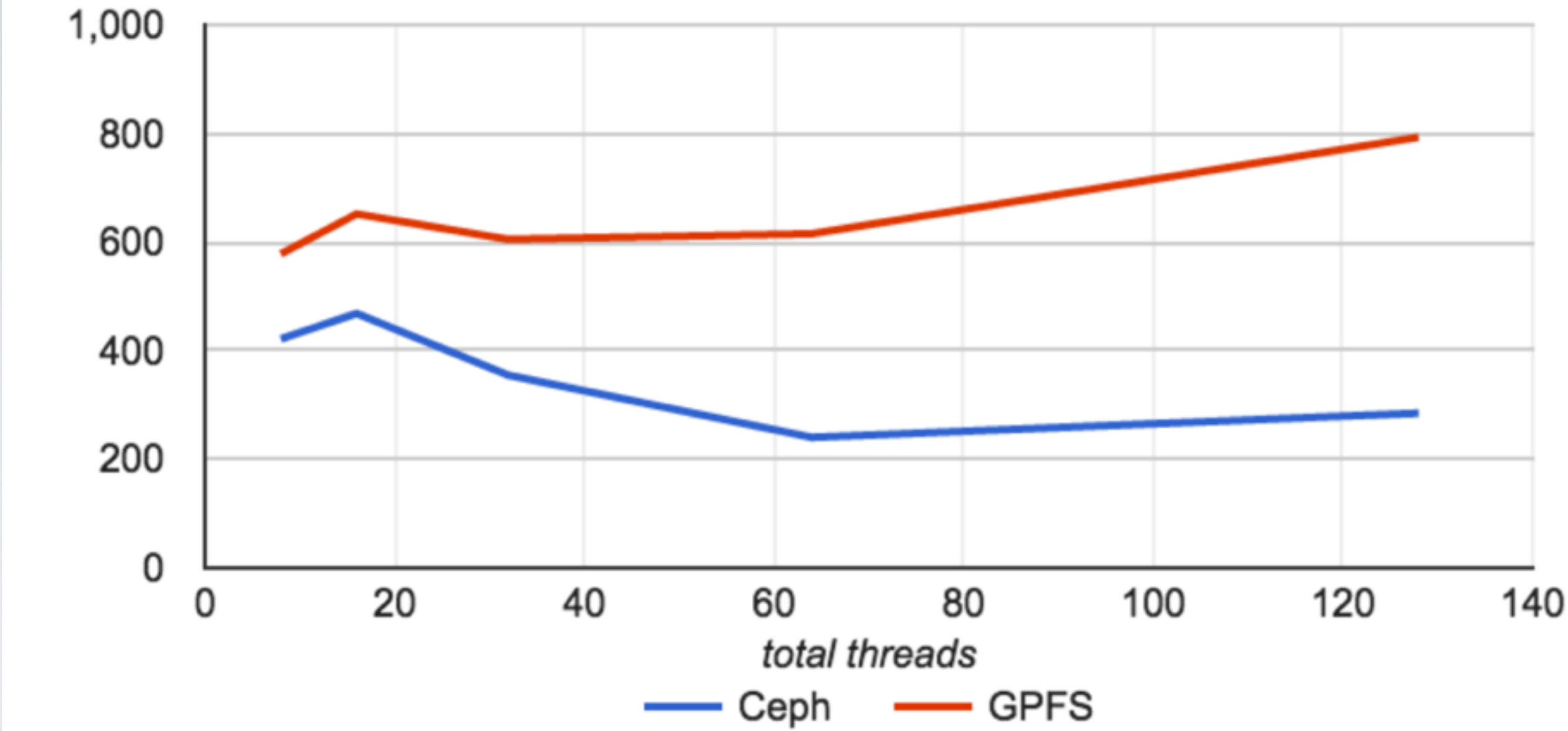
Ceph faster for <64 threads
(caching effect? lots of RAM)

GPFS up to 2x faster than Ceph for >64 threads

No intermediate storage of data at Seafile back-end while download?

UPLOAD RESULTS [MB/S]

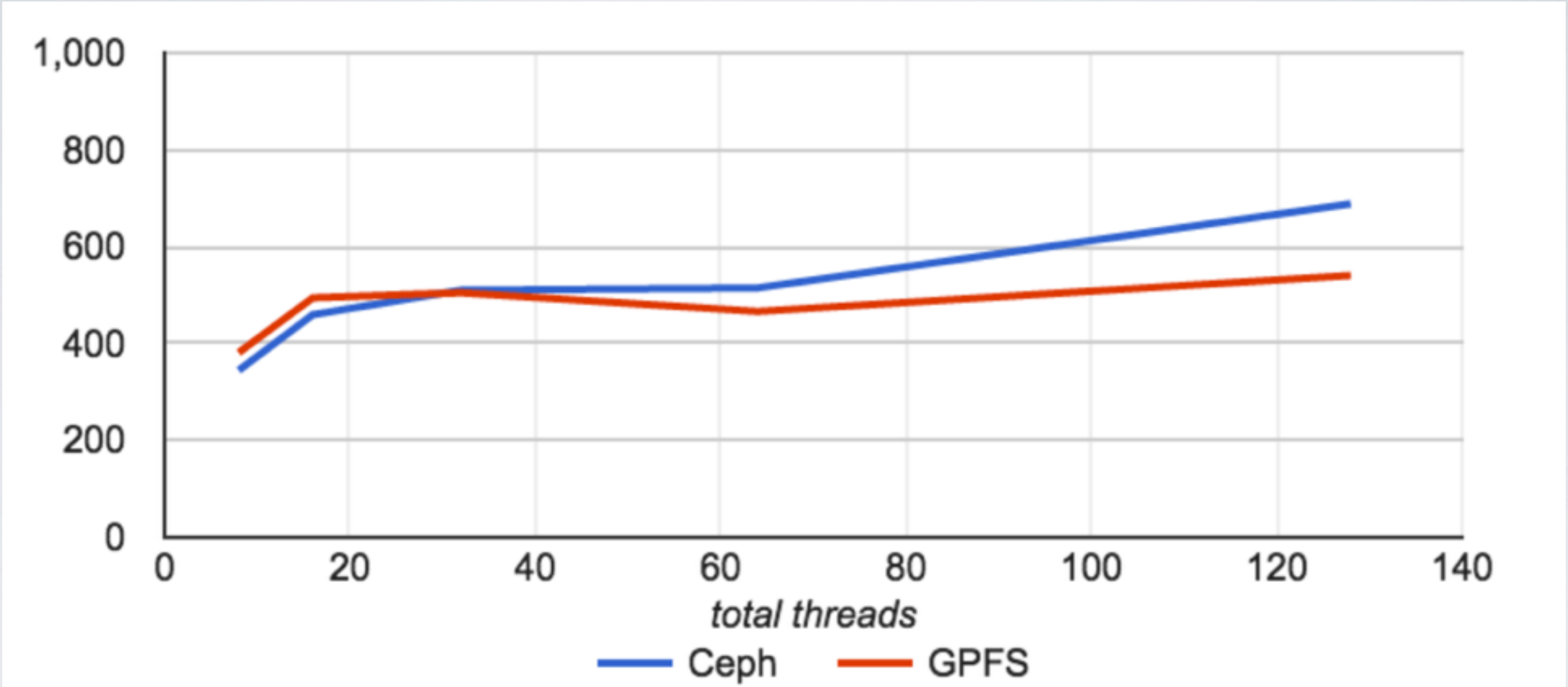
LARGE FILES TEST (4,4GB FILES)



GPFS is up to 3x faster than Ceph for large files
3x replication in Ceph?

DOWNLOAD RESULTS [MB/S]

LARGE FILES (4,4GB FILES)



GPFS performance is comparable to Ceph
(differences within 10%)

OUR APPROACH

- **BOX** is a **country-wide sync&share service** by PSNC:
  - large user base: not only based on a single institution
    - millions of files served
- We **applied BOX to the IMMERSIFY** use-case:
  - use a public instance of the service
  - and a Seafile client tools: incl. web, desktop and **drive**

WHAT IS SEADRIVE:

- **Virtual filesystem client:**
  - synchronises on-demand
    only these data that are accessed by the user
  - data ‚cached' on the user system
    and the used as local
  - Similar to project Infinity of Drobpox
- *As of now no other on-premise sync&share solution can make it*

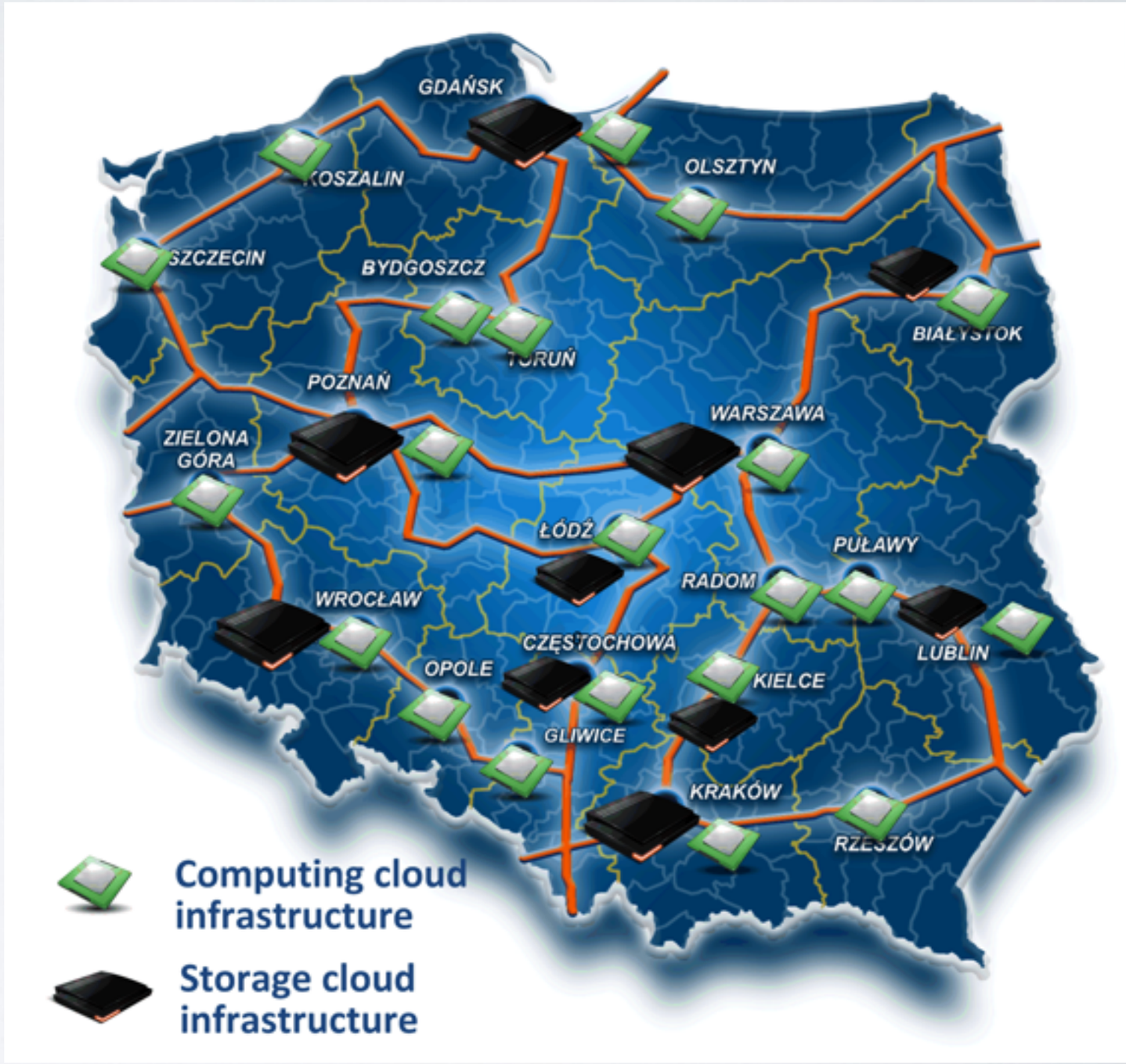
WHY SEADRIVE FOR IMMERSIFY?

- **Ease of use:**
  - **hides the complexity** of the workflow
    (these many files to be exchange)
  - **eliminates need** for **copying the data** manually / explicitely
    from PSNC to the Spin Digital site
  - provides **good integration** with other clients: Web, desktop
- **Robustness:**
  - Seafile will „stubbornly" synchronise the files down to the client
- **Performance:**
  - overall Seafile performance proven in our laboratory tests

PSNC & BOX: who we are, why we are doing this

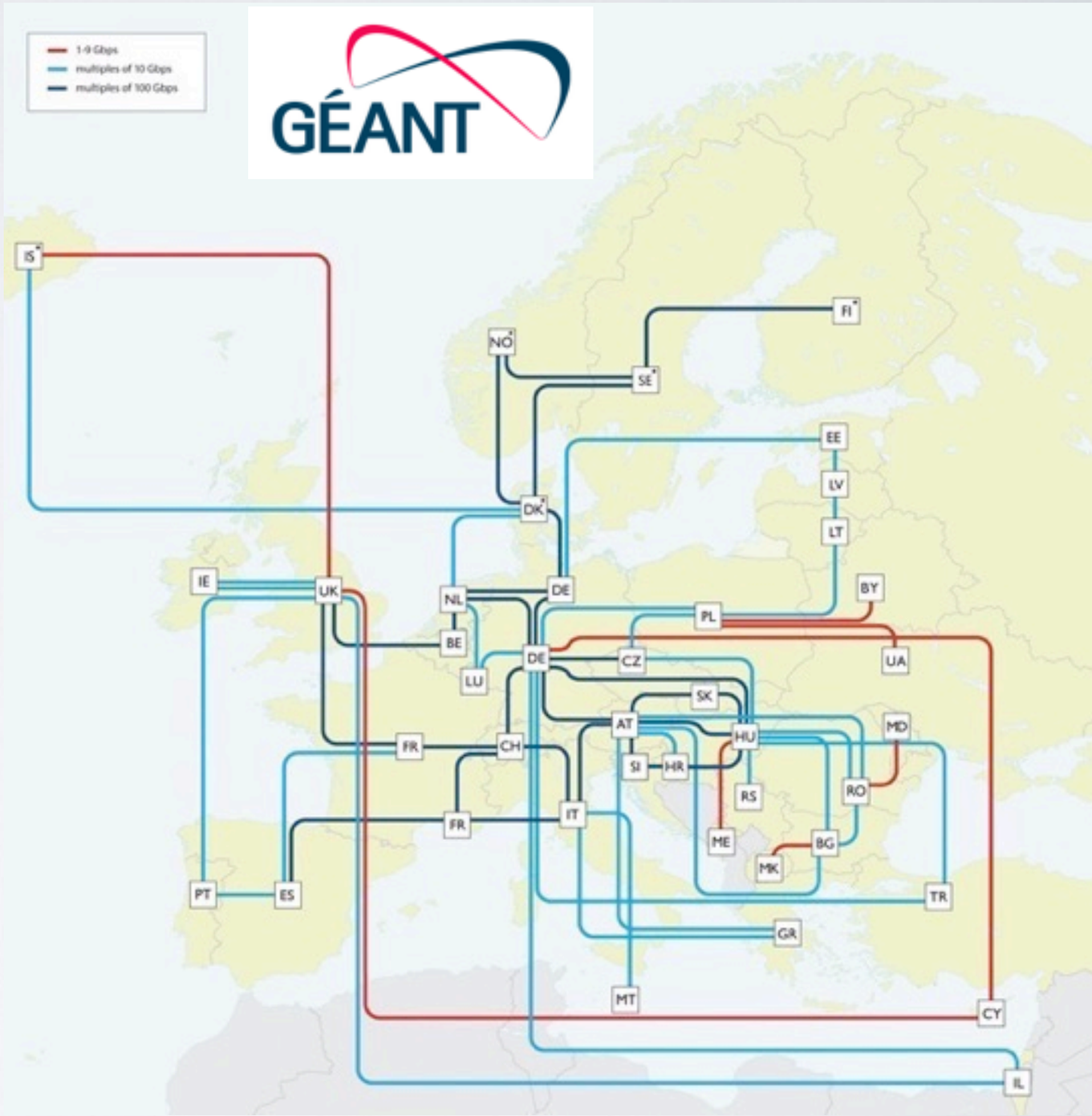
- POLISH NREN & SERVICES PROVIDER

- **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 THE EU ACADEMIC NETWORKS

- **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



OBSERVATIONS

FIRST BATTLE-FIELD EXPERIENCE

- **Seafile + Seadrive** is better than NFS server:
  - works using the Web protocols, no firewall passes
  - better - more fine-grained access control and authorisation
- **Throughput is OK, latency…:**
  - **Throughput:** we can sustain 10 Gib/s link with massive files
  - **Latency:** OK for codes (local buffer helps), not OK for interactive players
- Overall **the workflow is very simplified**
  - We use data ‚as-is' through whatever client: drive, web, desktop
  - Spin Digital can access ad-hoc any arbitrary dataset
  - Content updates or new content is propagated automatically

FUTURE WORK

- **Perform more synthetic benchmarks**
  - Basic tools such as iozone, fio (filesystem interface)
  - Build 4k video / coding process specifics tools or use codecs as the benchmark
  - Analyse latency and throughput + the efficiency of sync & share algorithm
- **Improve configuration**
  - TCP/IP tunning
  - Tunning Seafile parameters
- **Increase the scale of the tests:**
  - More sites perhaps
  - Longer and shorter distance (now it's ~280km Poznań-Berlin)

HIGH-LEVEL OBSERVATION

- **We believe that running services on premises still makes sense**
  - The functionality software available to us
    makes it possible to ‚compete' with public cloud services
    (Seafile's Seadrive vs Project Infinity of Dropbox)
  - Performance achieved can't be possible
    reached using public clouds
  - Budget-wise, using public clouds could be unaffordable
  - **We as NRENs and nerds :)**
    and thus **we have potential and willingness**
    to **work with users** at the case-by-case basis

EOF ;)

THANK YOU ;)

PSNC