SYNC&SHARE SOLUTION FOR MASSIVE MULTIMEDIA DATA

Maciej Brzeźniak, Krzysztof Wadowka, Paweł Woszuk
HPC Department

Maciej Głowiak, Maciej Stróżyk
New Media Department
AGENDA

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

1. Develop advanced video compression technology tailored for the needs of the VR video enabling delivering and displaying 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. Allow 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 workflow
  • 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**
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)

<table>
<thead>
<tr>
<th>SPEED</th>
<th>Seafile [GB/s]</th>
<th>theOther [GB/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x1GB file upload</td>
<td>0.17</td>
<td>0.11</td>
</tr>
<tr>
<td>5x1GB file download</td>
<td>0.29</td>
<td>0.71</td>
</tr>
</tbody>
</table>

LARGE FILES *)
- 5 GB file
# Seafile Performance

## Small Files*) Performance Test (2016)

<table>
<thead>
<tr>
<th>SPEED</th>
<th>Seafile [files-dirs/s]</th>
<th>theOther [files-dirs/s]</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client 1: upload</td>
<td>627</td>
<td>27</td>
<td>23x</td>
</tr>
<tr>
<td>Client 2: download:</td>
<td>940</td>
<td>43</td>
<td>22x</td>
</tr>
</tbody>
</table>

**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)**

<table>
<thead>
<tr>
<th>test</th>
<th>2016 test</th>
<th>2017, clustered Seafile 100kB files</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>single Seafile server; very small files - Linux kernel source</td>
<td>clustered Seafile 100kB files</td>
</tr>
<tr>
<td><strong>SPEED</strong></td>
<td>Seafie [files-dirs/s]</td>
<td>theOther [files-dirs/s]</td>
</tr>
<tr>
<td><strong>Client 1:</strong> upload</td>
<td>627</td>
<td>27</td>
</tr>
<tr>
<td><strong>Client 2:</strong> download</td>
<td>940</td>
<td>43</td>
</tr>
</tbody>
</table>
Having paid IBM already for GPFS, use them for sync & share?

Use Ceph as everybody does ;)

- Seafile server
- GPFS
- NSD client
- $$$$

- Seafile server
- NFS client

- Seafile server
- libRados client

- Seafile server
- RADOS

- GPFS

- GPFS

- Ceph
**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?
GPFS is up to 3x faster than Ceph for large files
3x replication in Ceph?
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 Dropbox
  • 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
• **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 Seafire 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