DSS

Data & Storage Services

CERN IT Department

CERN Cloud Storage Evaluation

Geoffray Adde, Dirk Duellmann, Maitane Zotes
CERN IT

HEPiX Fall 2012 Workshop

October 15-19, 2012
Institute of High Energy Physics, Beijing, China

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Outline

CERN IT Department

- What are the interests in cloud storage implementations and protocols?
- How can we insure they can also be realised in the HEP environment?
- Test plans for two S3 implementations
  - OpenStack/Swift
  - Openlab collaboration with Huawei
- Results after first testing phase
  - important contributions from Lu Wang (IHEP)
- Test plans for upcoming period

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

What cloud storage ?

CERN IT Department

- Storage used by jobs running in a computational cloud
  - network latency impacts what is usable depending on type of applications
- Storage build from (computational) cloud node resources
  - storage life time = life time of node
- Storage service by commercial (or private) providers exploiting similar scaling concepts as computational clouds
  - clustered storage with remote access
  - with cloud protocol and modified storage semantic
- Term may be valid for all of the above
  - but here I will refer to the last group of storage solutions

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Why Cloud Storage & S3 Protocol?

- Cloud computing and storage gain rapidly in popularity
  - Both as private infrastructure and as commercial service
  - Several investigations are taking place also in the HEP and broader science community
- Price point of commercial offerings may not (yet?) be comparable with services we run at CERN or WLCG sites, but
  - Changes in semantics, protocols, deployment model promise increased scalability with reduced deployment complexity (TCO)
  - Market is growing rapidly and we need to understand if promises can be confirmed with HEP work loads
  - Need to understand how cloud storage will integrate with (or change) current HEP computing models

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

S3 Semantics and Protocol

CERN IT Department

- Simple Storage Service (Amazon S3)
  - "just" a storage service
    - in contrast to eg Hadoop, which comes with a distributed computation model exploiting data locality
      - (Hadoop also being evaluated in IT-DSS - but not reported in here)
  - uses a language independent REST API
    - http(s) for transport
- Provide additional scalability by
  - focussing on a defined subset of posix functionality
  - partitioning of namespace into independent buckets
- S3 protocol alone does not provide scalability
  - eg if added naively on top of a traditional storage system
  - scalability gains to be proven for each S3 implementation

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN Interest in Cloud Storage

- Main Interest: Scalability and TCO
  - can we run cloud storage systems to complement or consolidate existing storage services?
- Focus: storage for physics and infrastructure
  - near-line archive pools, analysis disk pools, home directories, virtual machine image storage
- Which areas of the storage phase space can be covered well?
- First steps:
  - setup and run a cloud storage service of PB scale
  - confirm scalability and/or deployment gains

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Potential Interest for WLCG

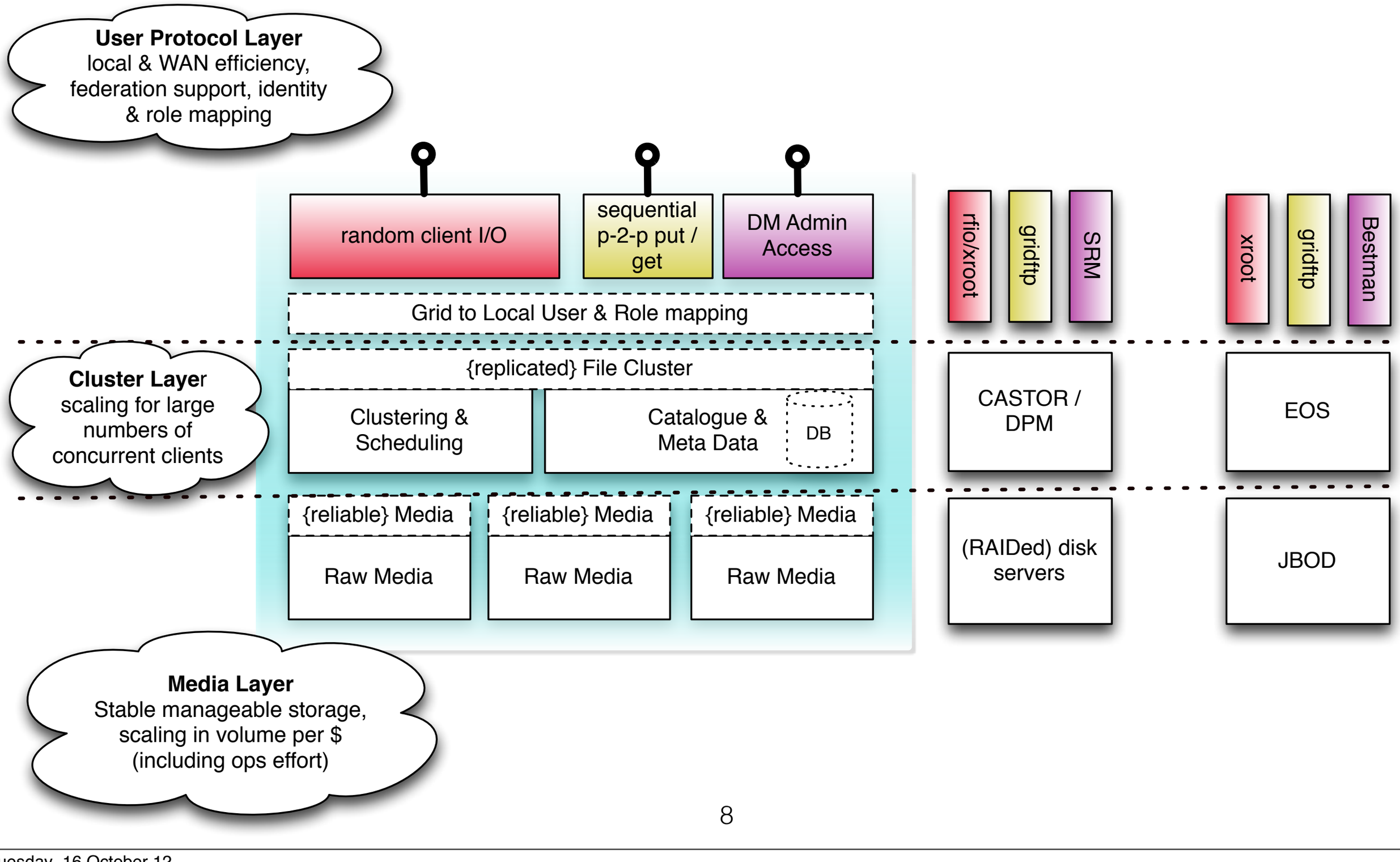
CERN IT Department

- S3 Protocol could be a standard interface for access, placement or federation of physics data
- Allowing to provide (or buy) storage services without change to user application
  - large sites may provide private clouds storage on acquired hardware
  - smaller sites may buy S3 or rent capacity on demand
- First Steps
  - successful deployment at one site (eg CERN)
  - demonstrate data distribution across sites (S3 implementations) according to experiment computing models

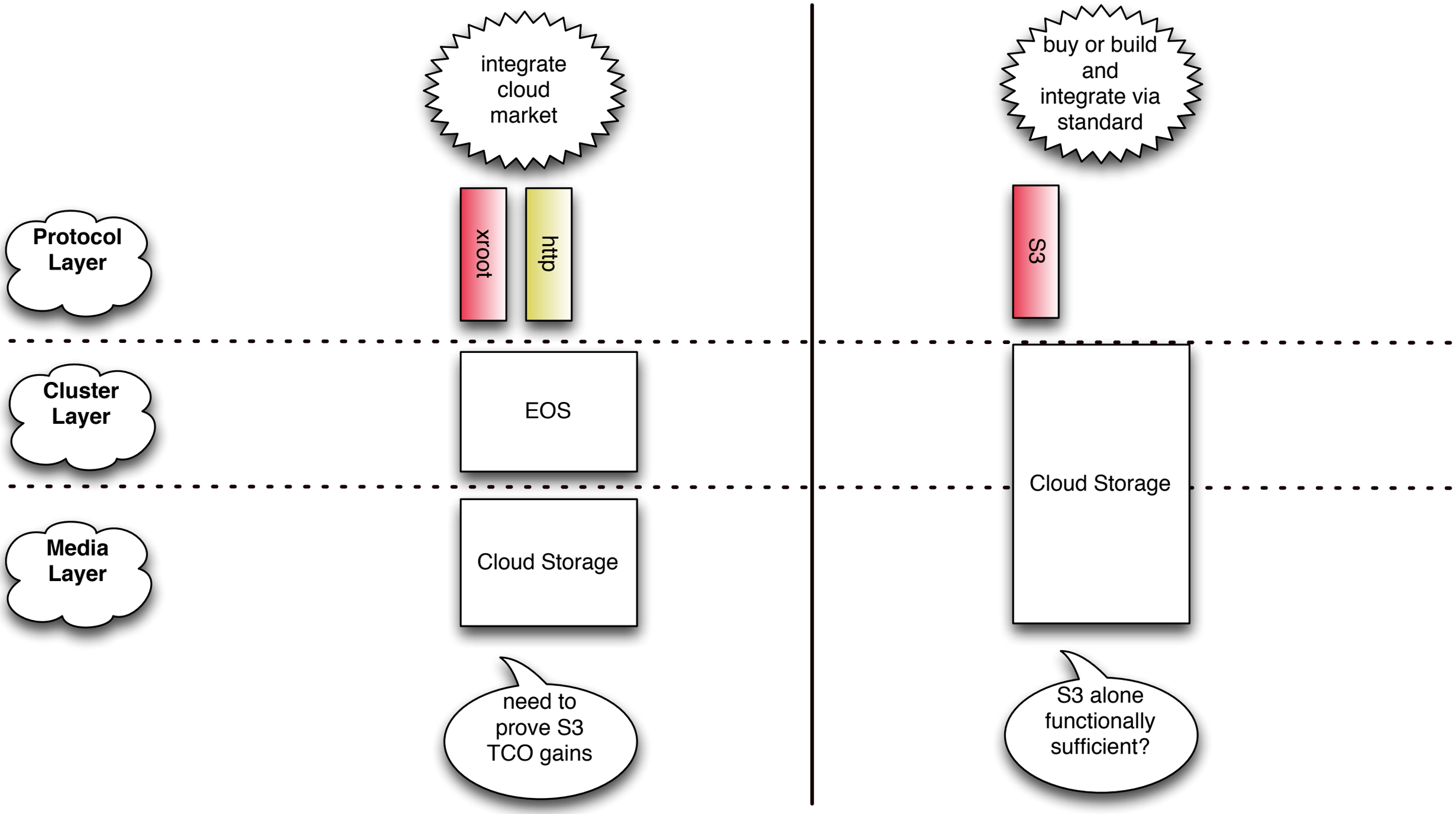
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

Component Layering in current SEs



Potential Future Scenarios



Common Work Items

CERN IT Department

- Common items for OpenStack/Swift and Huawei systems

☑ Define the main I/O pattern of interest

- ☐ based on measured I/O patterns in current storage services (eg archive, analysis pool, home dir, grid home?)

☑ Define implement and test a S3 functionality test

- ☑ define S3 API areas of main importance
- ☑ develop a S3 stress / scalability test
  - ☐ scale up to several hundred concurrent clients (planned for August)
  - ☑ copy-local and remote access scenarios

- Define key operational use cases and classify human effort and resulting service unavailability
  - ☑ add remove disk servers (incl. draining)
  - ☑ h/w intervention by vendor (does not apply to Huawei)
  - ☑ s/w upgrade
  - ☐ power outage

☐ Compare performance and TCO metrics to existing services at CERN

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Work Items - Huawei Appliance

CERN IT Department

- ☑ Support commissioning of Huawei storage system
  - 0.8 PB system in place at CERN
- ☑ Share CERN test suite and results with Huawei
  - Tests (including ROOT based ones) are regular being run by Huawei development team
- ☑ Perform continuous functional and performance tests to validate new Huawei s/w releases
  - ☑ maintain a list of unresolved operational or performance issues
- Schedule and execute large scale stress test
  - ☑ S3 test suite
  - ☐ Hammercloud with experiment applications

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

OpenStack/SWIFT

- Actively participate in fixing of CERN issues with released software
  - ☑ build up internal knowledge about SWIFT implementation and defects
  - ☑ probe our ability to contribute and influence the release content from the OpenStack foundation

☑ Run the same functionality and stress tests as for the Huawei system

☐ Visit larger scale SWIFT deployments to get in-depth knowledge about

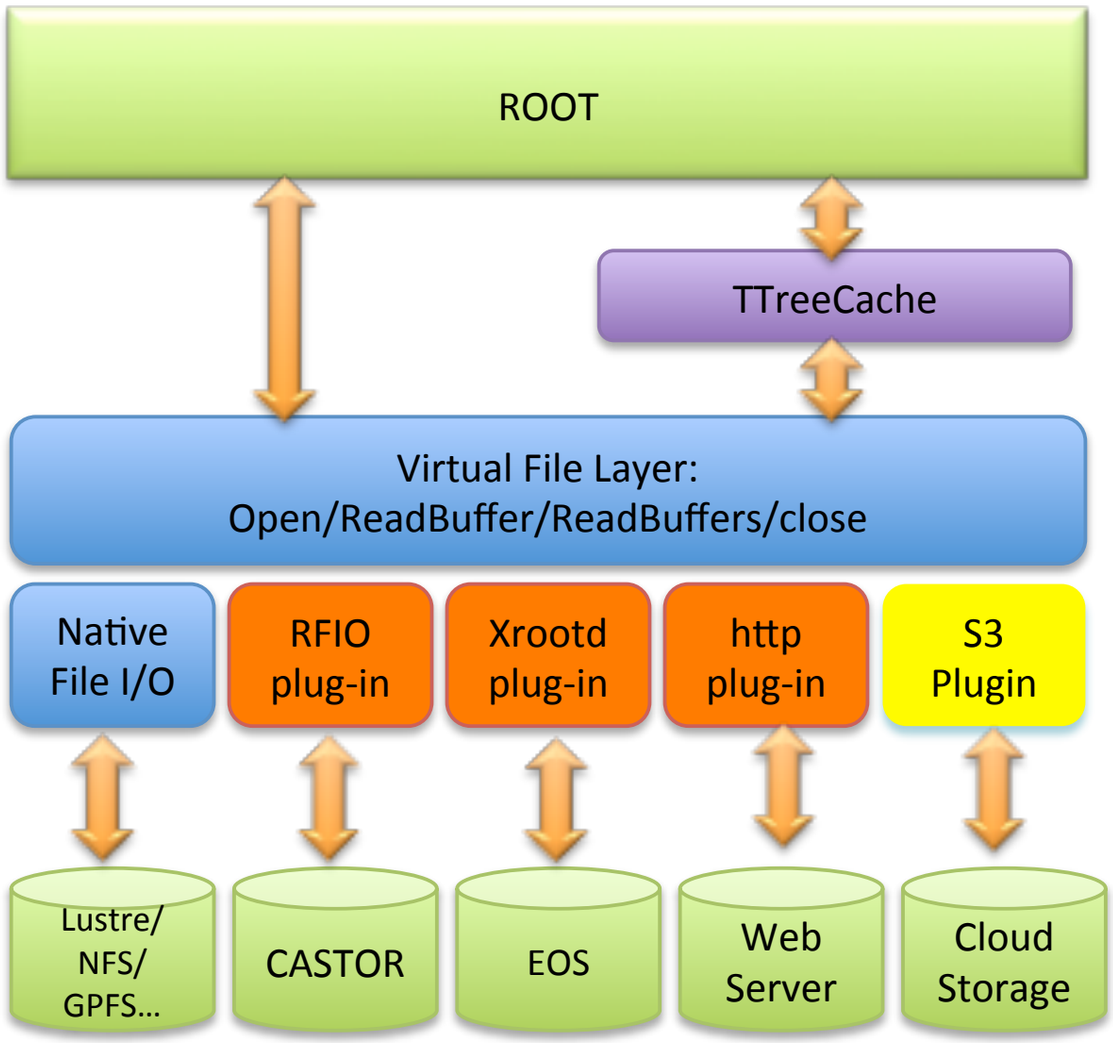
- level of overlap between open software and in-house deployed additions / improvements
- compare I/O pattern in typical deployments with CERN services

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

S3 plug-in for ROOT

CERN IT Department

- Based on the http plug-in
- Adapts to S3 protocol
- Supports *vector read* requests issued by TTree cache
  - Huawei added multi-range read to S3 implementation
- Integrated with the distributed I/O test framework



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Fuse-based file system

CERN IT Department

- Open source prototype S3FS
  - http://code.google.com/p/s3fs/

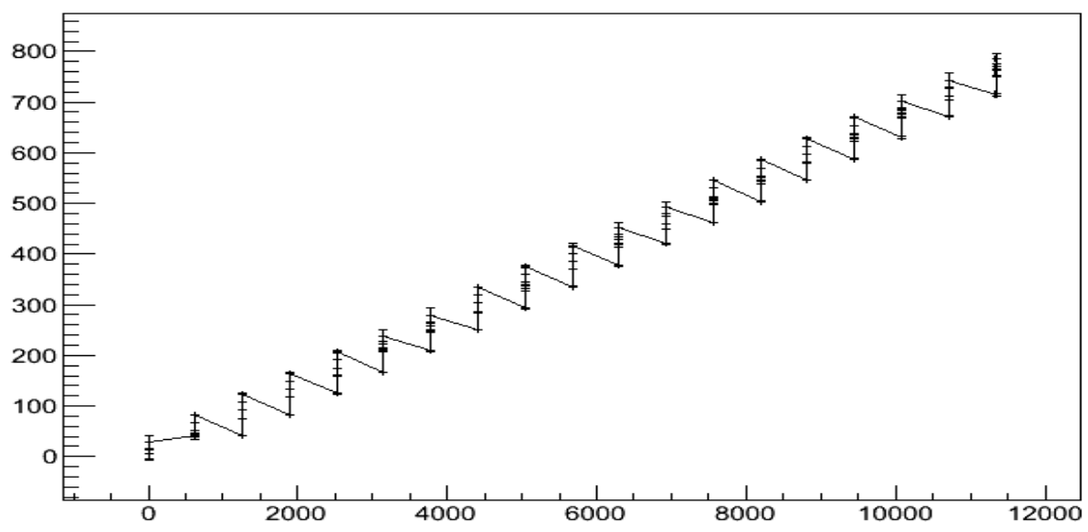
- Current limitations:
  - Can only mount one single bucket instead of the whole system
  - Instead of remote I/O, file is downloaded to local cache during “open”
  - “df” returns not relevant information

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

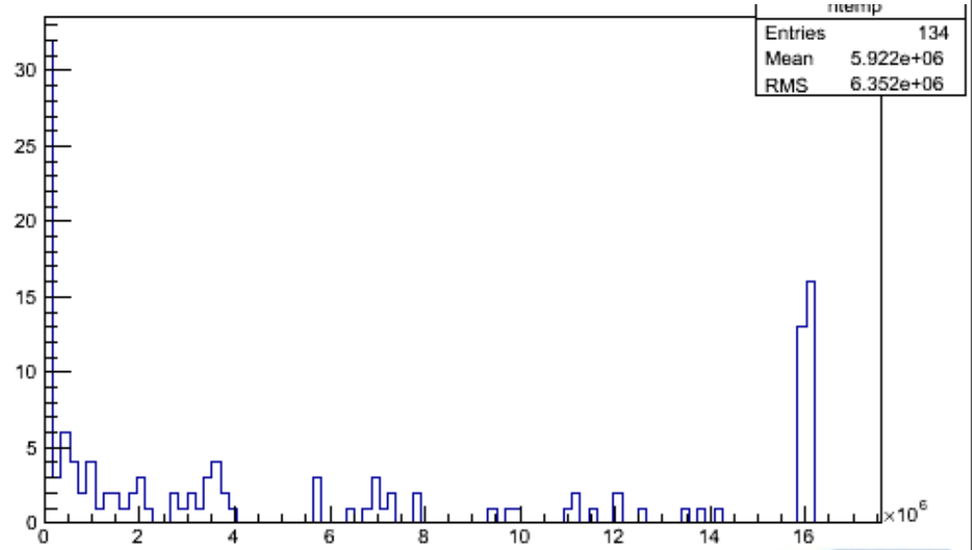
Test Job

CERN IT Department

- A real ATLAS ROOT file
  - 793MB on disk, 2.11 GB after decompression
  - 11918 entries, 5860 branches ,cache size=30MB
- Reads in entries sequentially in "*physics tree*"



Reading Offset vs Entry

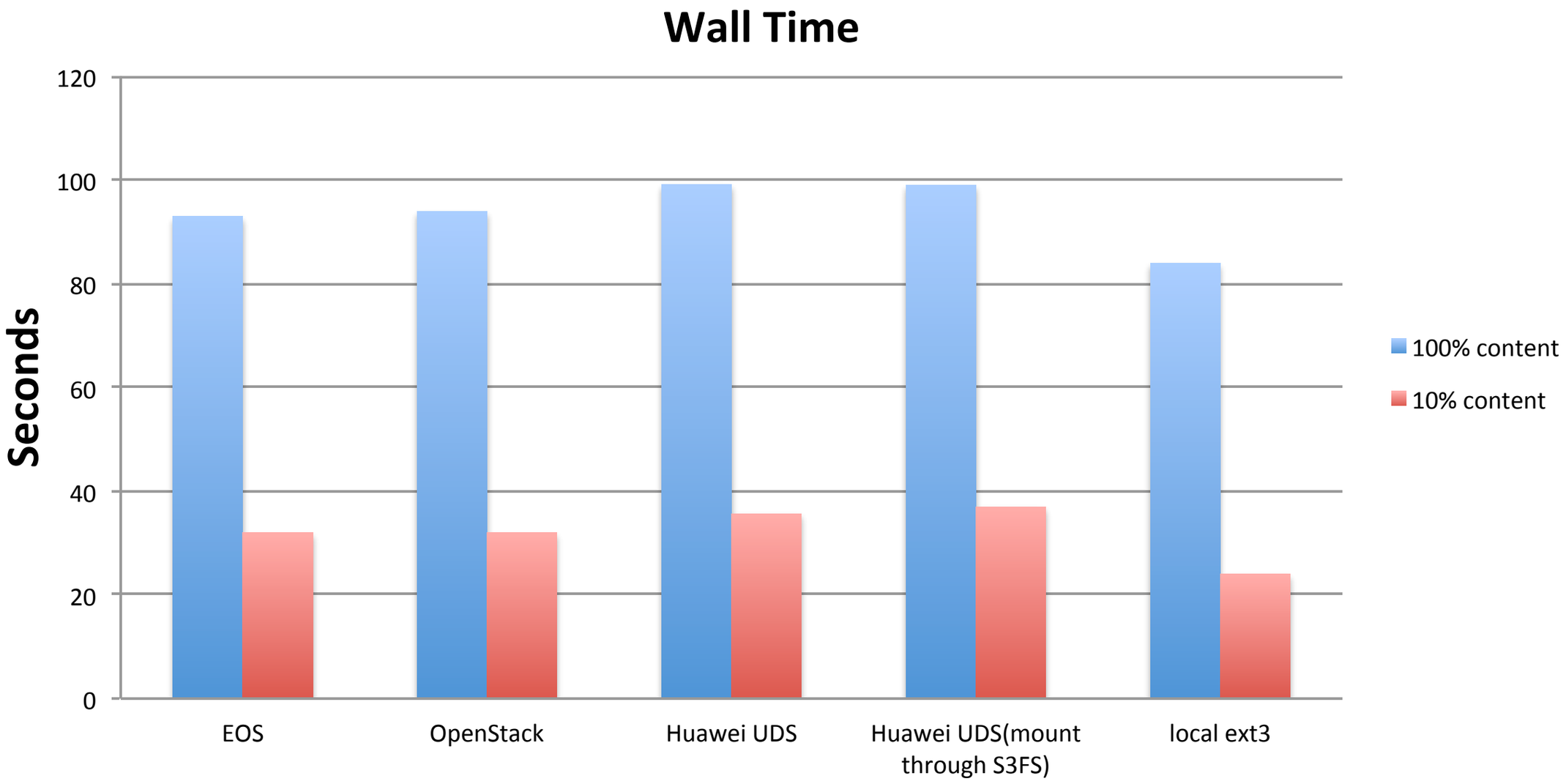


Distribution of read Size

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Test Result

CERN IT Department



- Single client performance
  - reach similar performance than local fs access
  - already with previous Huawei release

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Preliminary Results

CERN IT Department

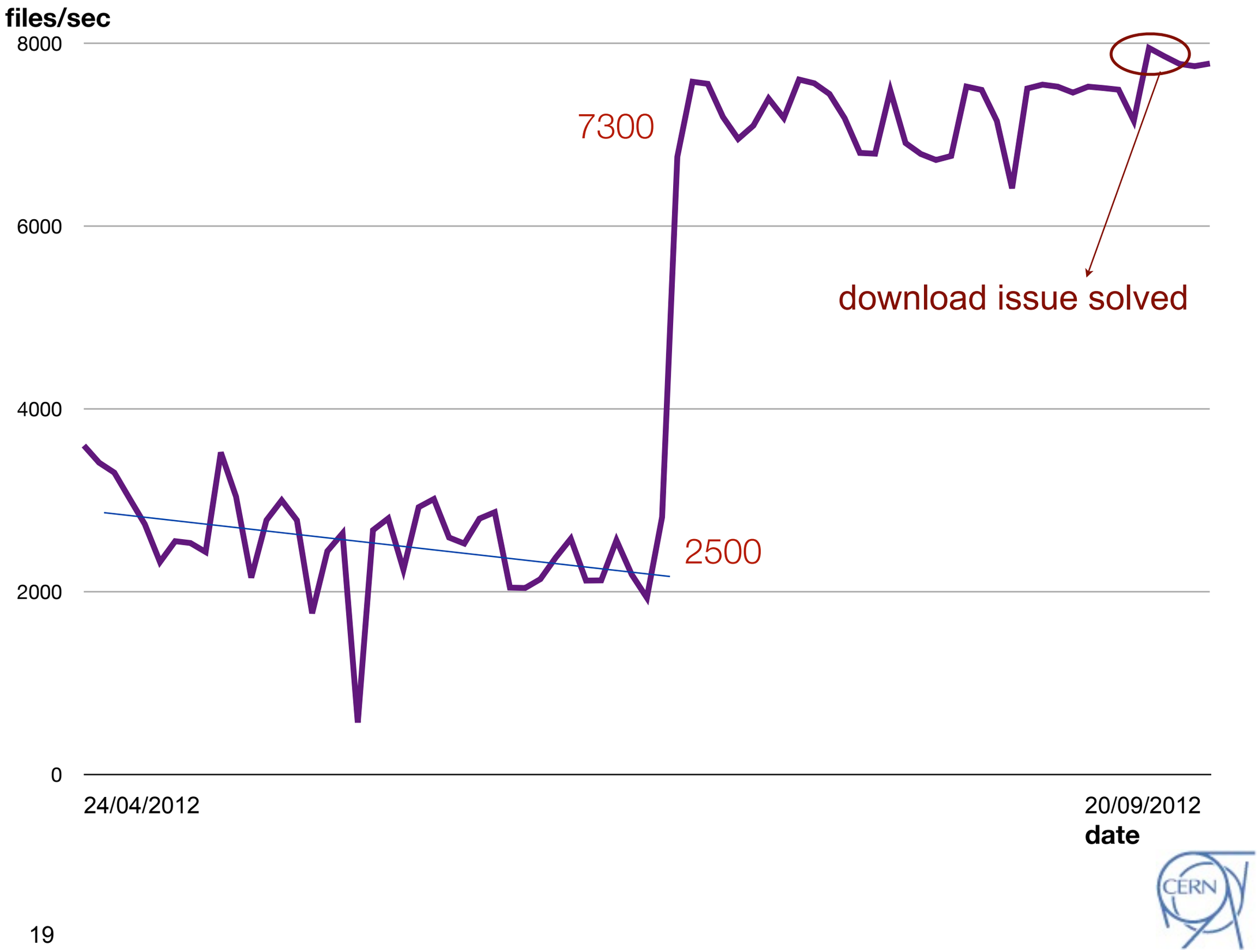
- OpenStack/Swift and Huawei reach similar performance as EOS or local filesystem
  - for full file access or 10% fraction of file
- Analysis type access using the ROOT S3 plugin
  - naive use (no TTreeCache) of both S3 implementations shows significant overhead
  - with enabled ROOT cache and vector read this overhead is removed
- S3 filesystem reaches performance of S3 plugin based access
  - assuming that local cache space (/tmp) is available
- No authentication and authorisation yet for S3 storage
  - not yet mapped from certificates used in WLCG

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CERN

“Morning test” - Long term stability

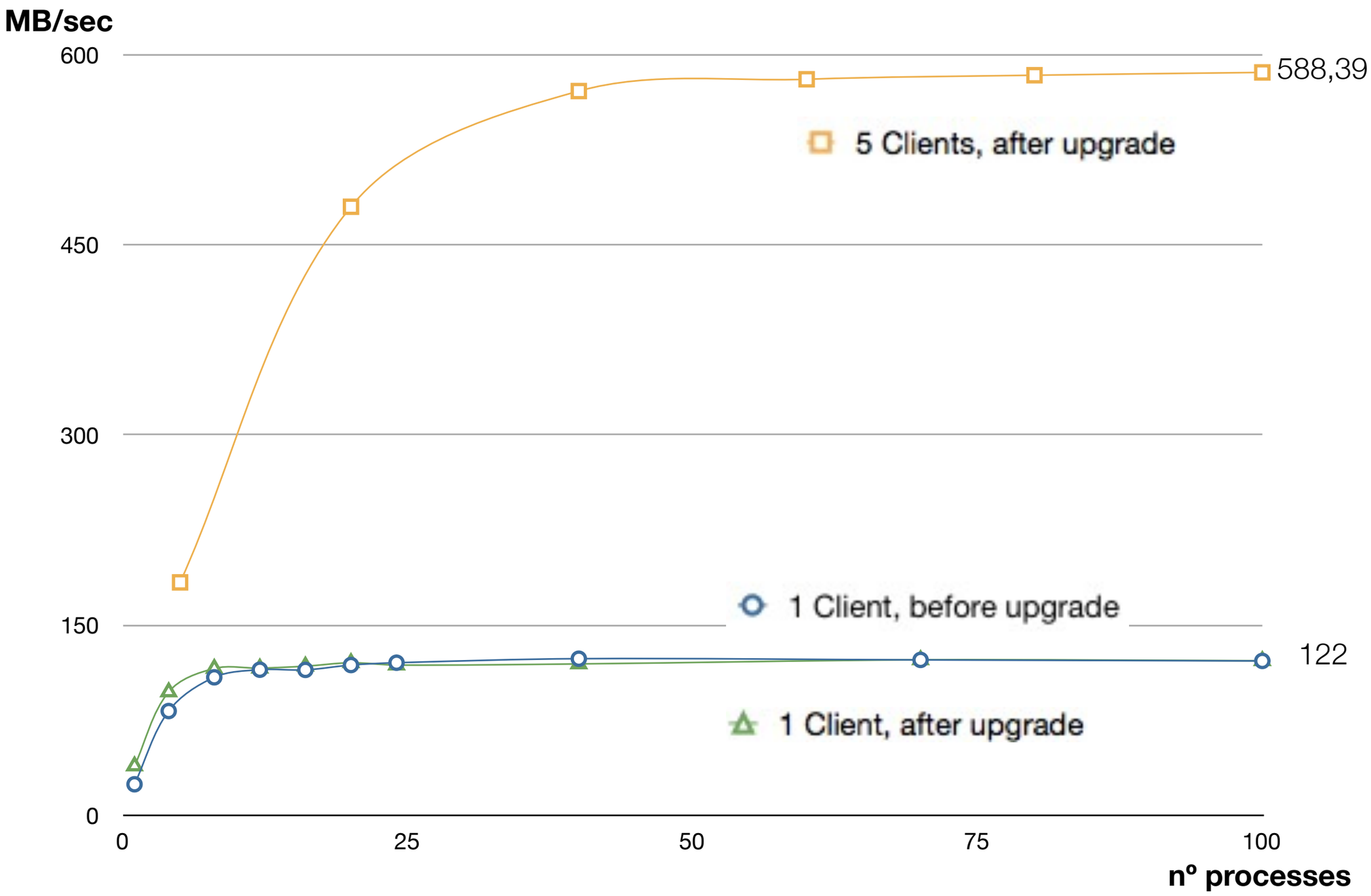
CERN IT Department



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

100MB Upload - Huawei

CERN IT Department

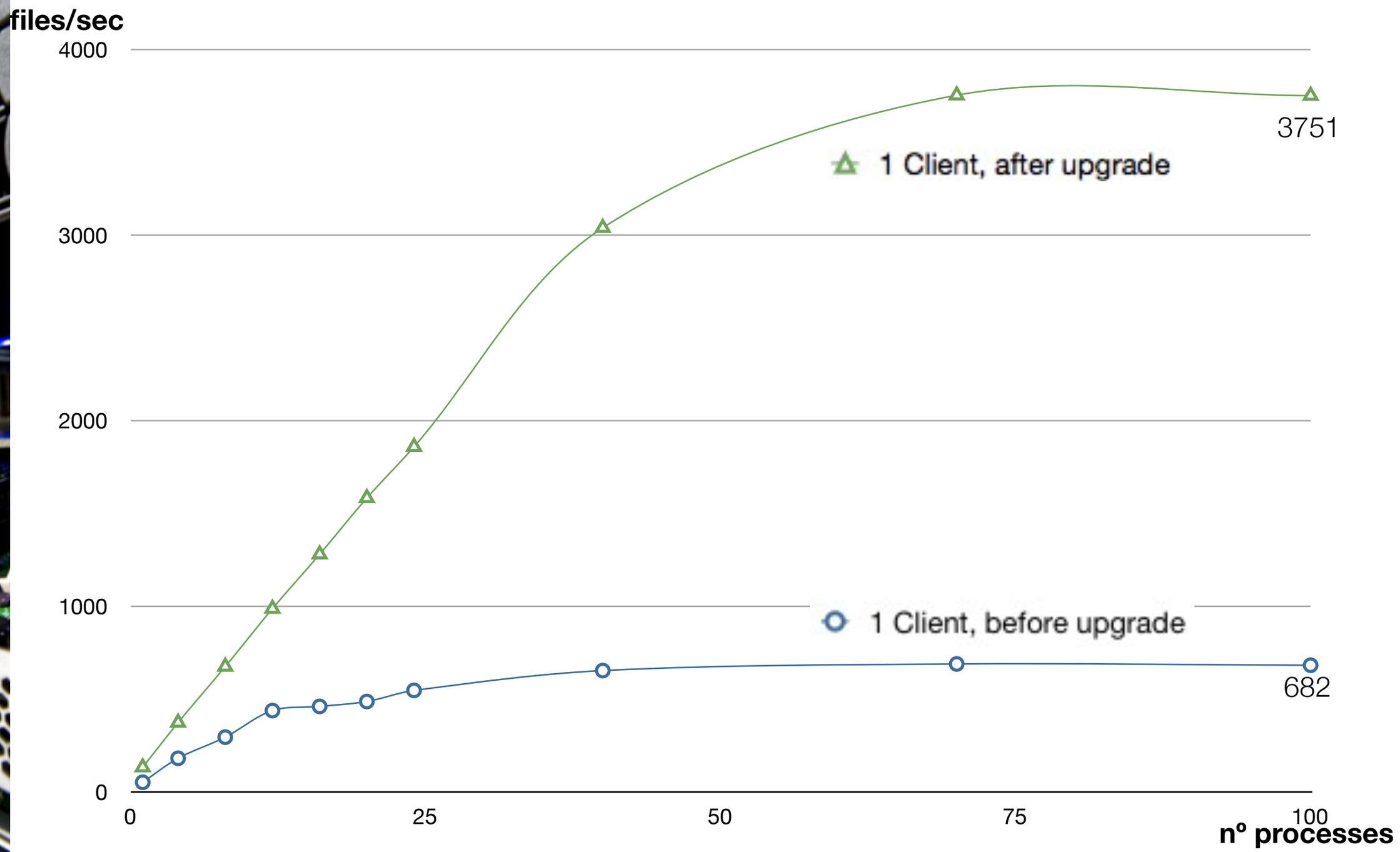


CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

No problem to reach bandwidth limit of 5Gb
(from 5 client boxes - up to 100 processes total)

4KB Download - Huawei

CERN IT Department

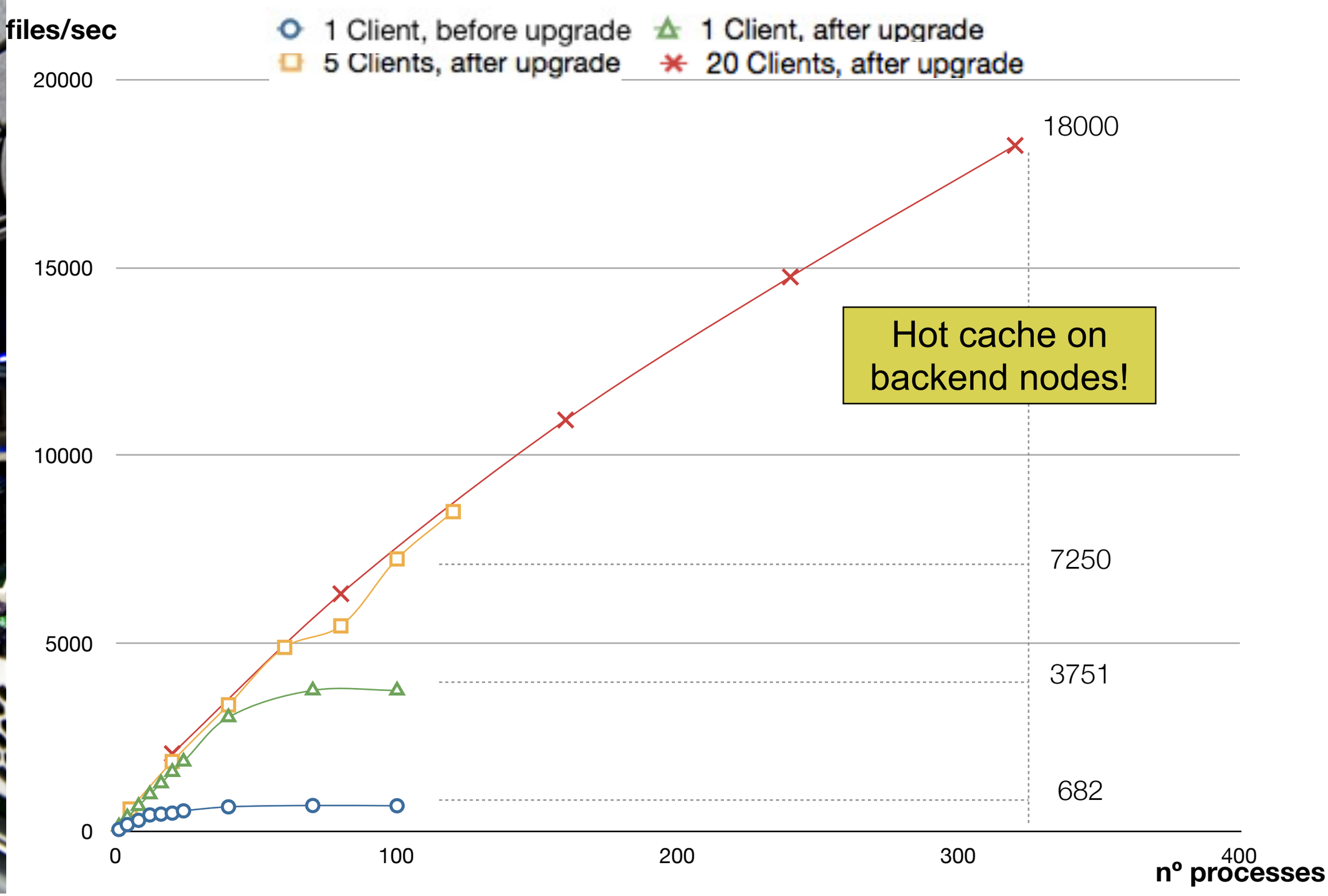


Performance after upgrade: five times better

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

DSS 4KB Download - Huawei

CERN IT Department

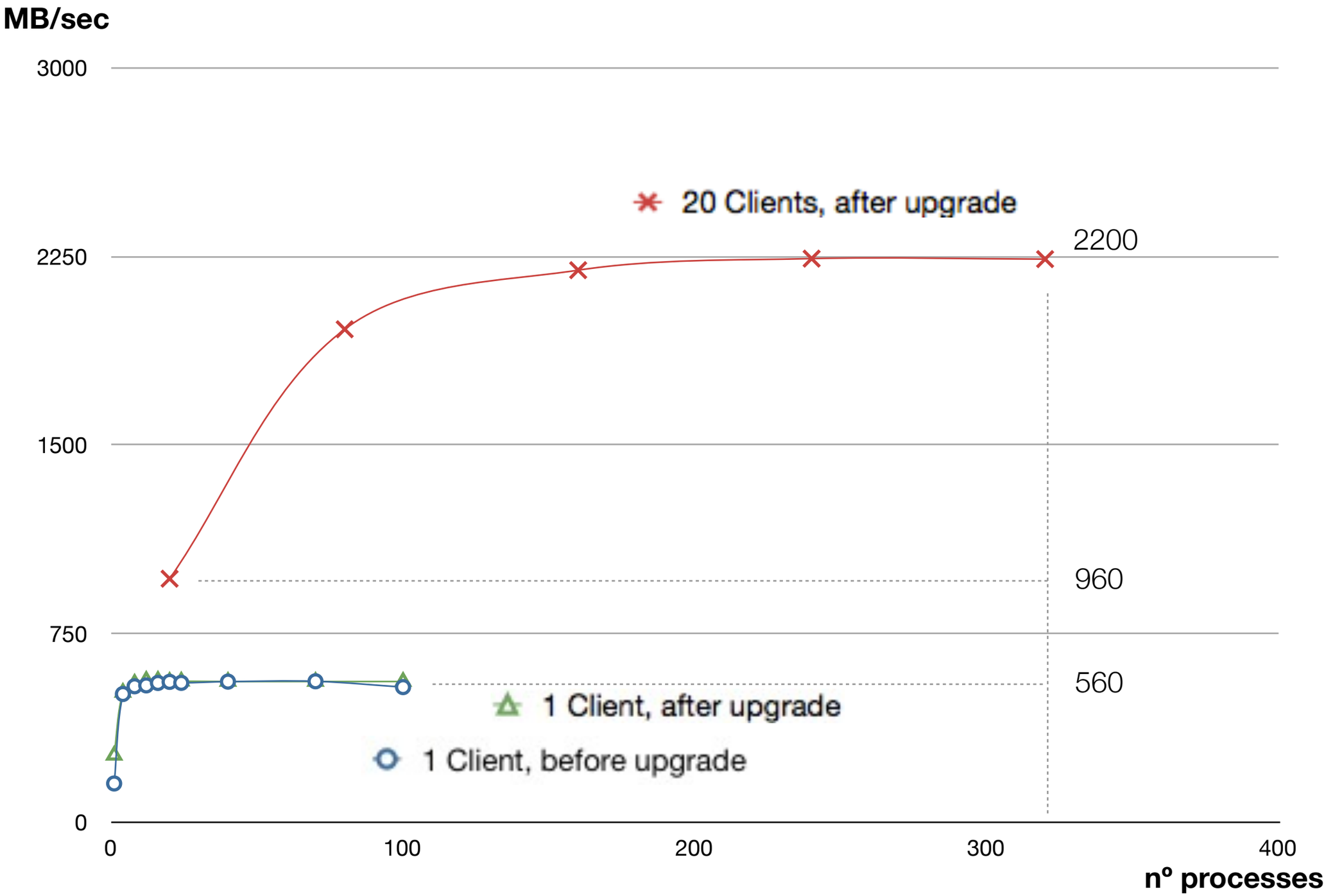


Close to linear metadata scaling (up to 300 procs)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

100MB Download - Huawei

CERN IT Department

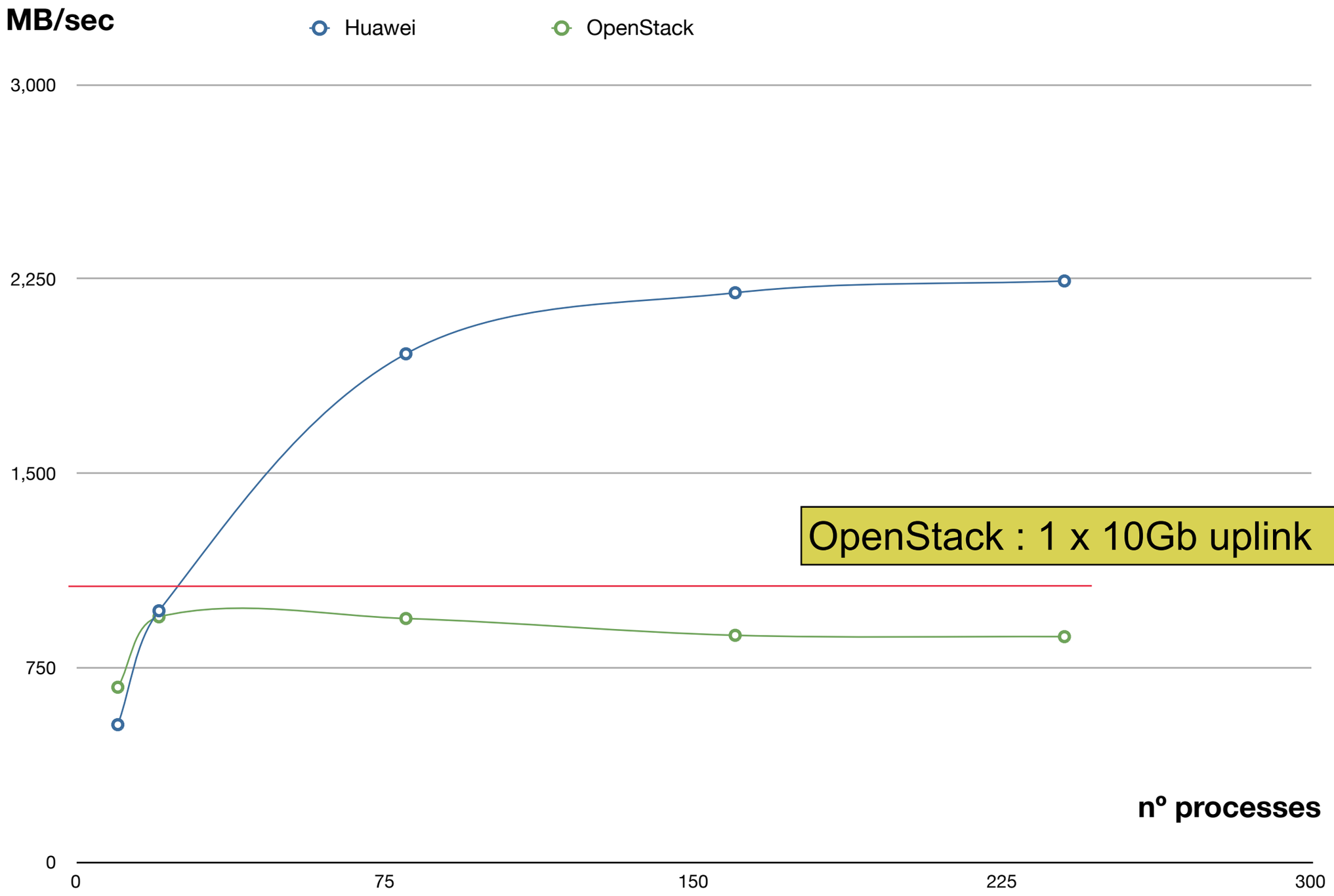


20 clients reach bandwidth limit of 18Gb

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

100MB Download - OpenStack

CERN IT Department



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Summary

- Cloud storage evaluation with two S3 implementations has started this year
  - Client performance of local S3 based storage looks comparable to current production systems
  - Achieved expected stability and aggregate performance (Huawei)
    - Metadata performance up to 8k files/s
      - proved expected scalability of the system
    - Total throughput up to 18 Gb/s
      - fully maxed out the 2 fibres available
      - balanced system with 350MB/s per OSC
    - Minor technical problems found and resolved rapidly
      - productive collaboration with Huawei in context of CERN openlab
  - OpenStack/Swift looks promising, but need to complete test suite
- Realistic TCO estimation can not yet be done in a small (1PB) test system w/o real users access

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Future plans

CERN IT Department

- 2012 - short term
  - Further increase scalability range with additional network and client resources
  - Analyse performance impact of cache(s) and journal
  - Collect feedback ATLAS workload management system
  - Exercise transparent upgrade procedure (Huawei)
  - Complete/compare OpenStack/Swift measurements
- 2013 - next year
  - Multiple datacenter tests (eg in collaboration with IHEP)
  - Erasure code impact on performance and space overhead
  - Prove transparent failure recovery with consumer disks
- Goal for 2013
  - Evaluate TCO gains of S3 storage as part of a production service
  - Candidate services being evaluated
    - CVMfs on S3 storage (interest from CVMfs team)
    - AFS on S3 storage (prototype developed by Rainer Többicke)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

- Thank you!

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it