Huawei Cloud Storage

Seppo S. Heikkila
Maria Arsuaga Rios
CERN IT

Openlab Major Review Meeting
13th of February 2014
CERN, Geneva

Introduction

Motivation

– Cloud storage market is growing fast

– CERN uses custom made storage solutions

Question

> “Are cloud storages able to meet the High Energy Physics (HEP) data storage requirements?”

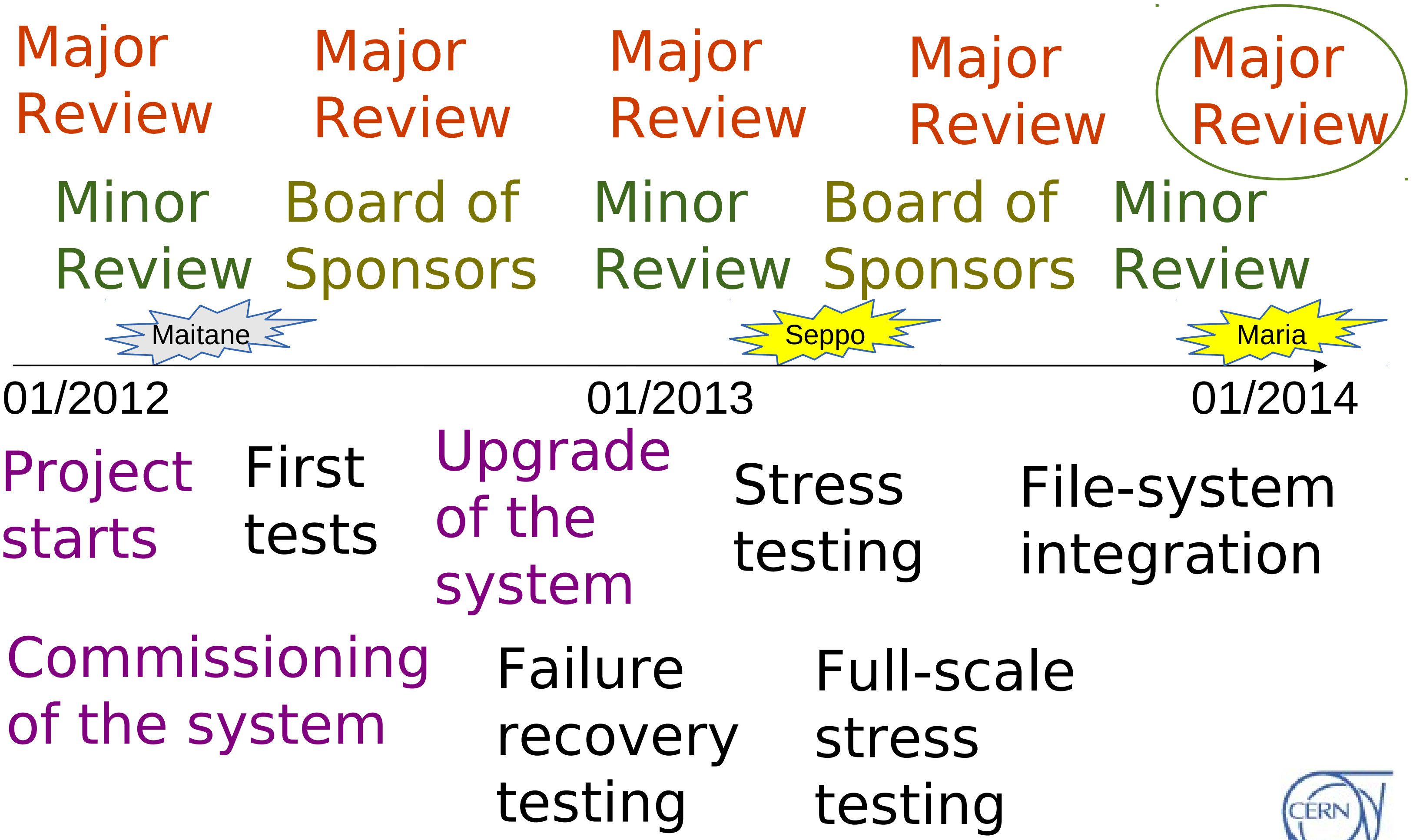
Method

– Evaluate scalability and fault-tolerance

– Test with real applications

DSS Timeline

2 years of Huawei...



Major Review

Major Review

Major Review

Major Review

Major Review

Minor Review

Board of Sponsors

Minor Review

Board of Sponsors

Minor Review

Project starts

First tests

Upgrade of the system

Stress testing

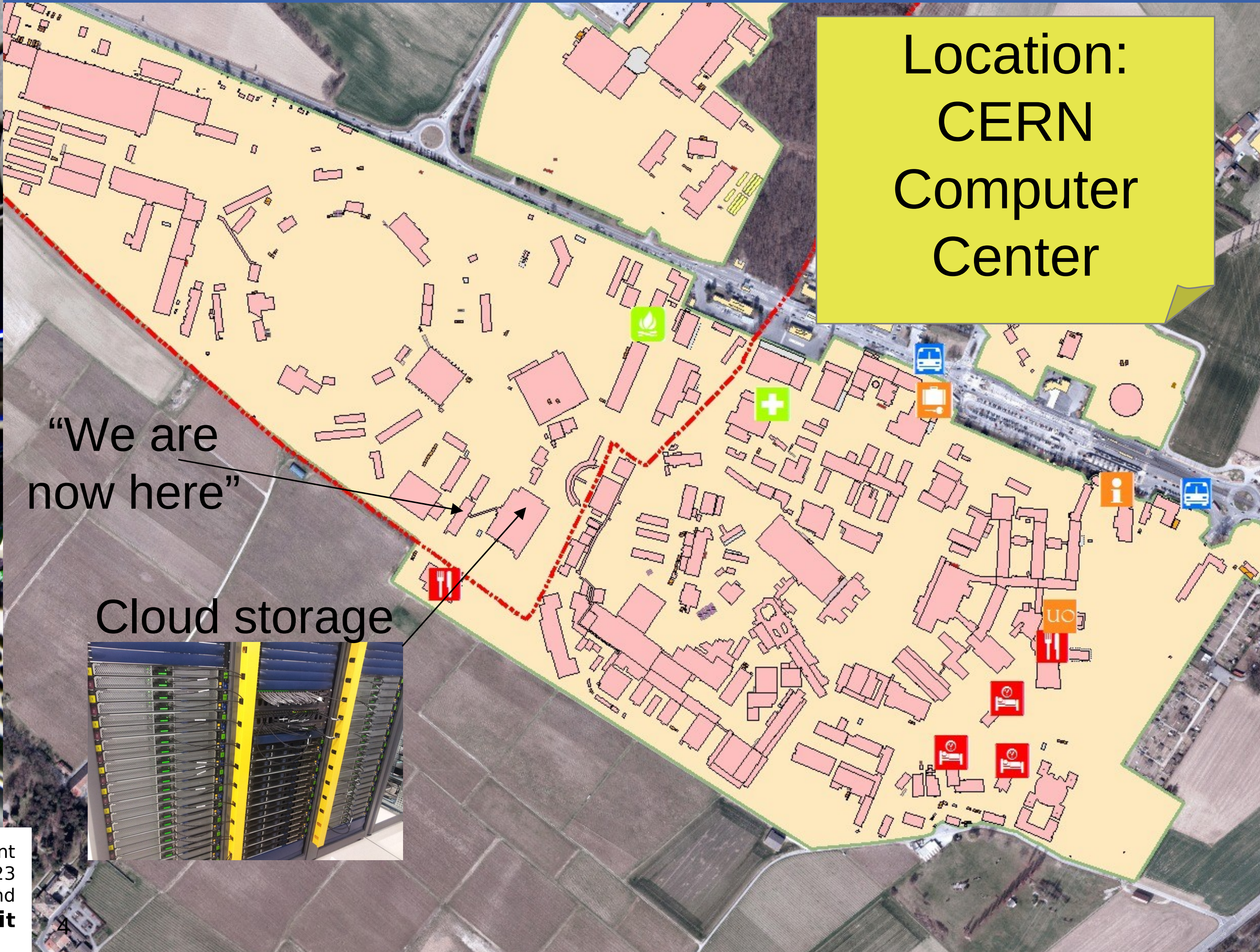
File-system integration

Commissioning of the system

Failure recovery testing

Full-scale stress testing

Huawei cloud storage

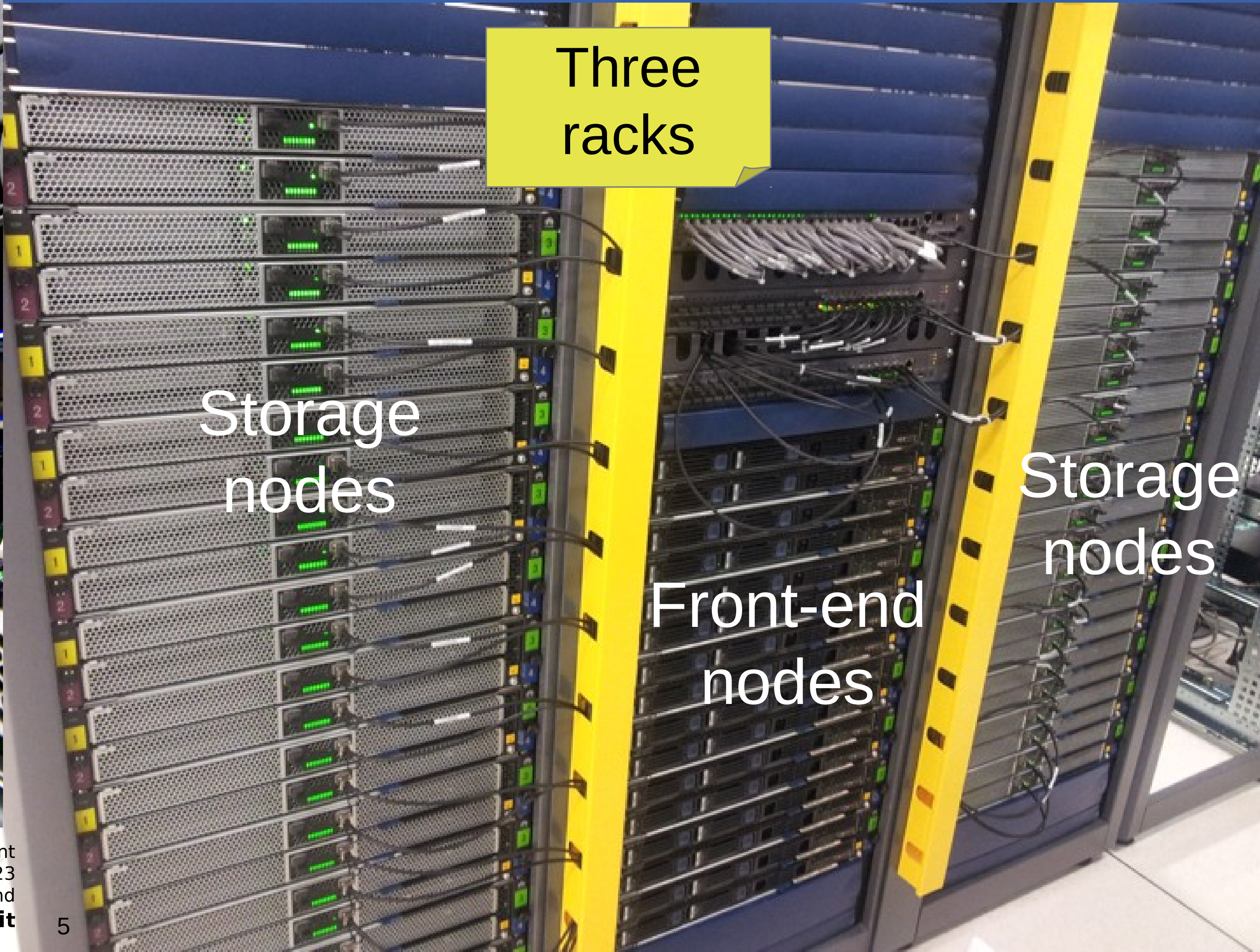


Location: CERN Computer Center

"We are now here"

Cloud storage

Huawei cloud storage setup



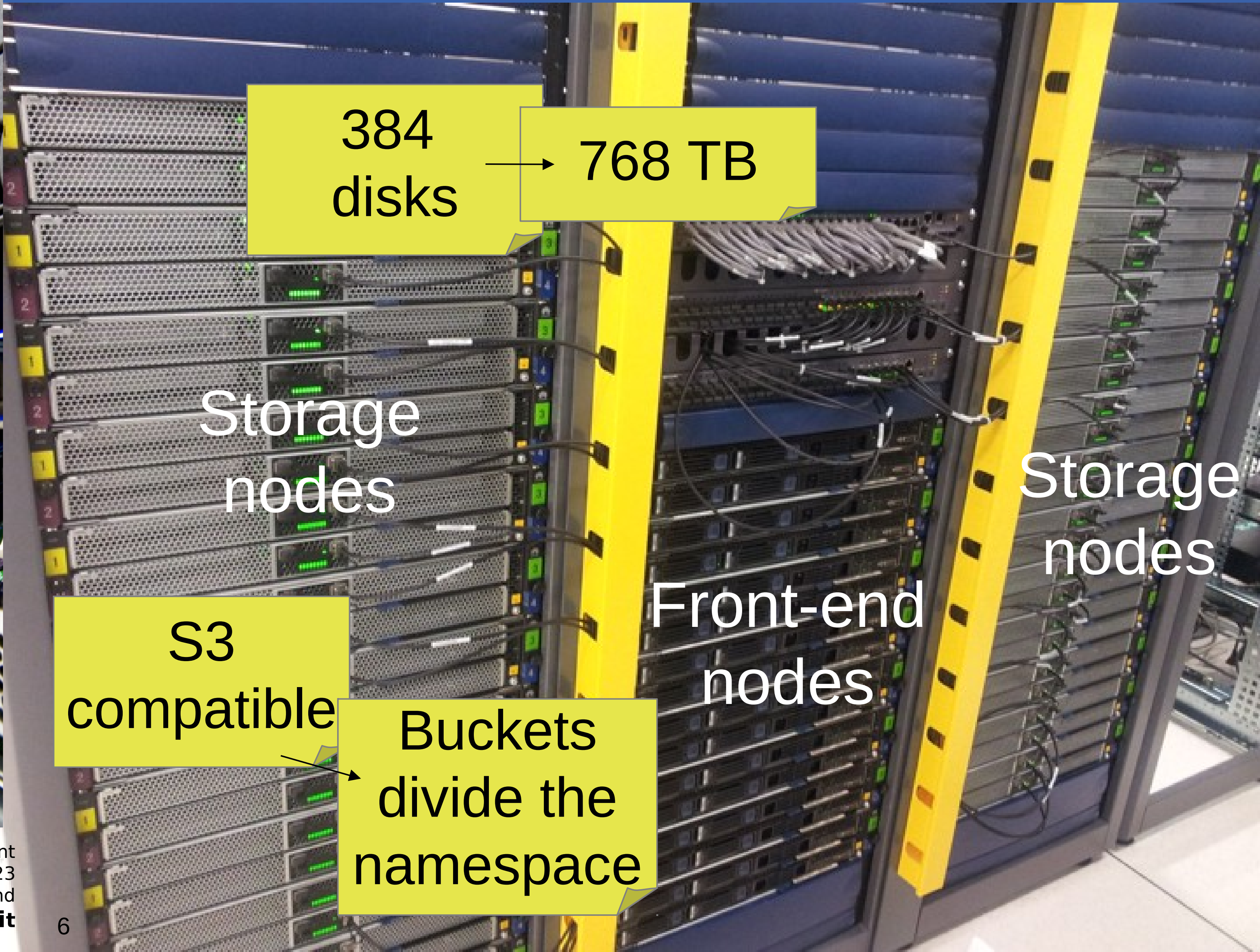
Three racks

Storage nodes

Front-end nodes

Storage nodes

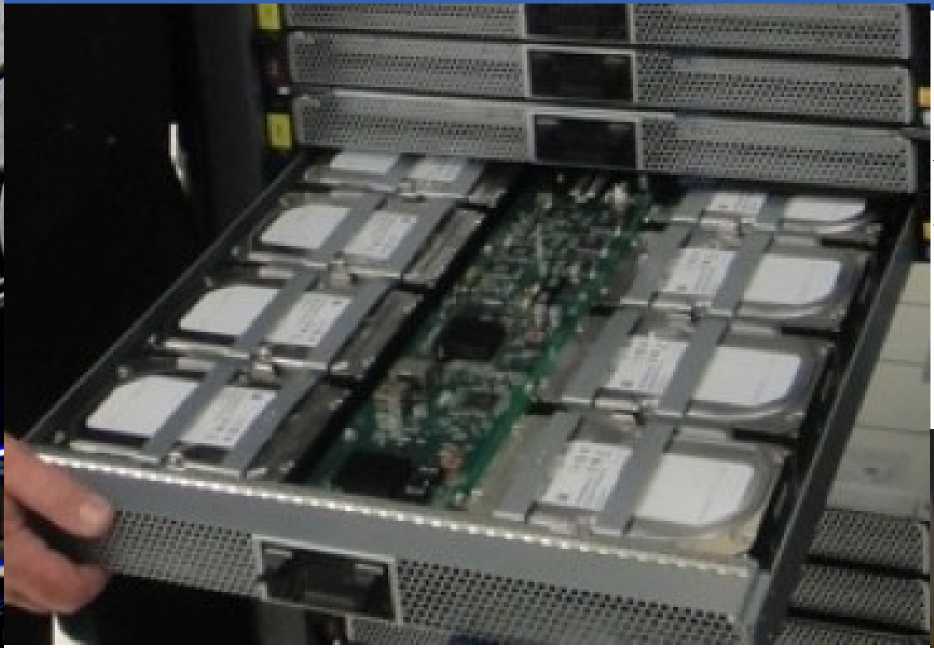
Huawei cloud storage setup



384 disks → 768 TB

S3 compatible → Buckets divide the namespace

Storage nodes



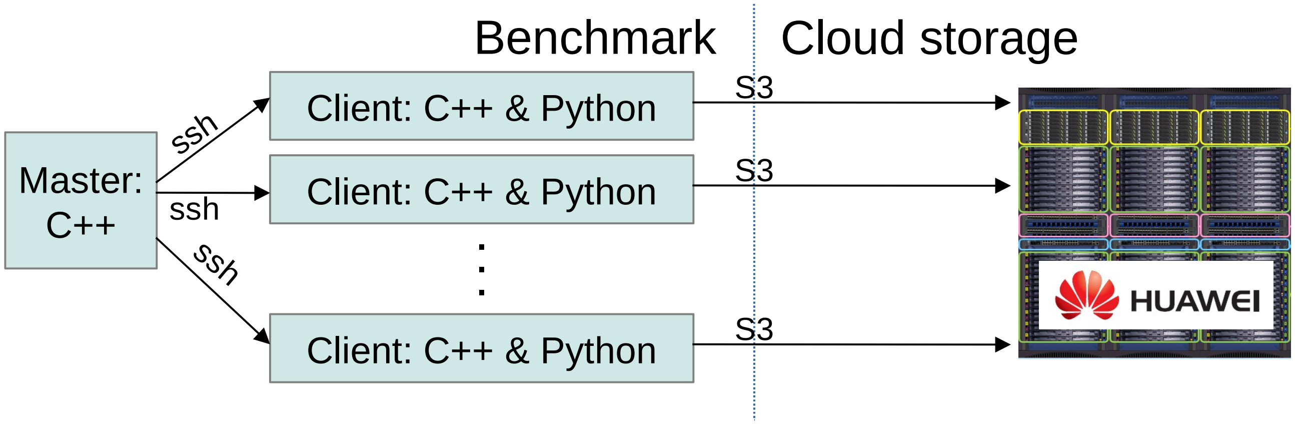
One chassis has two blades

Each blade has eight storage nodes

Benchmark setup

Distributed C++ benchmark

- Integrated with ROOT
- Client nodes connected with ssh
- S3 Python library to read and write files
- Histograms about specific metrics
  - Operation time, read/write speed, CPU/memory utilisation



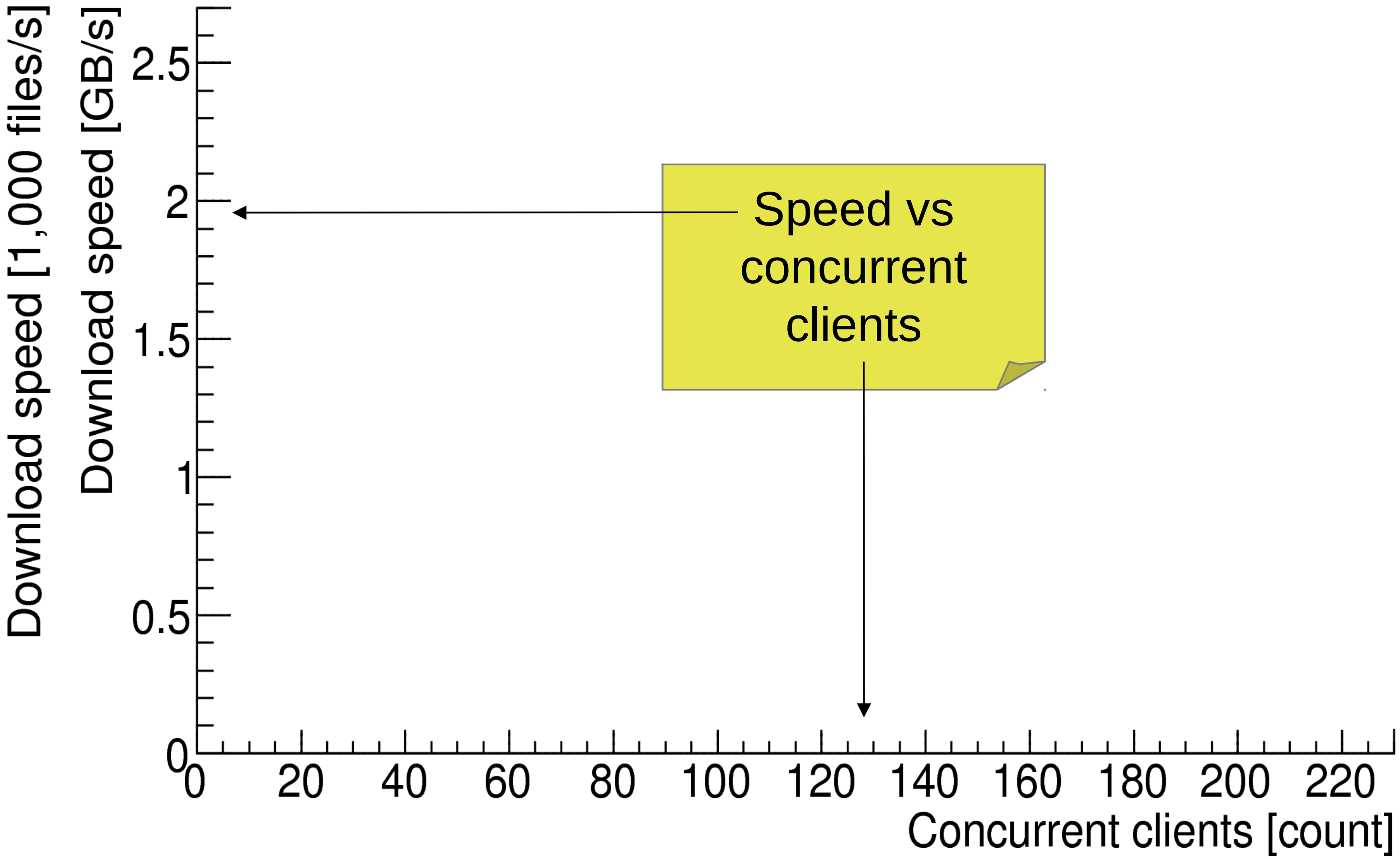
CHEP conference paper

"Zotes Resines M, Heikkila SS, Duelmann D, Adde G, Toebbicke R, Hughes J and Wang L. Evaluation of the Huawei UDS cloud storage system for CERN specific data, International Conference on Computing in High Energy and Nuclear Physics (CHEP) 2013, Amsterdam, The Netherlands, 14 October 2013"

Presented and submitted in October!

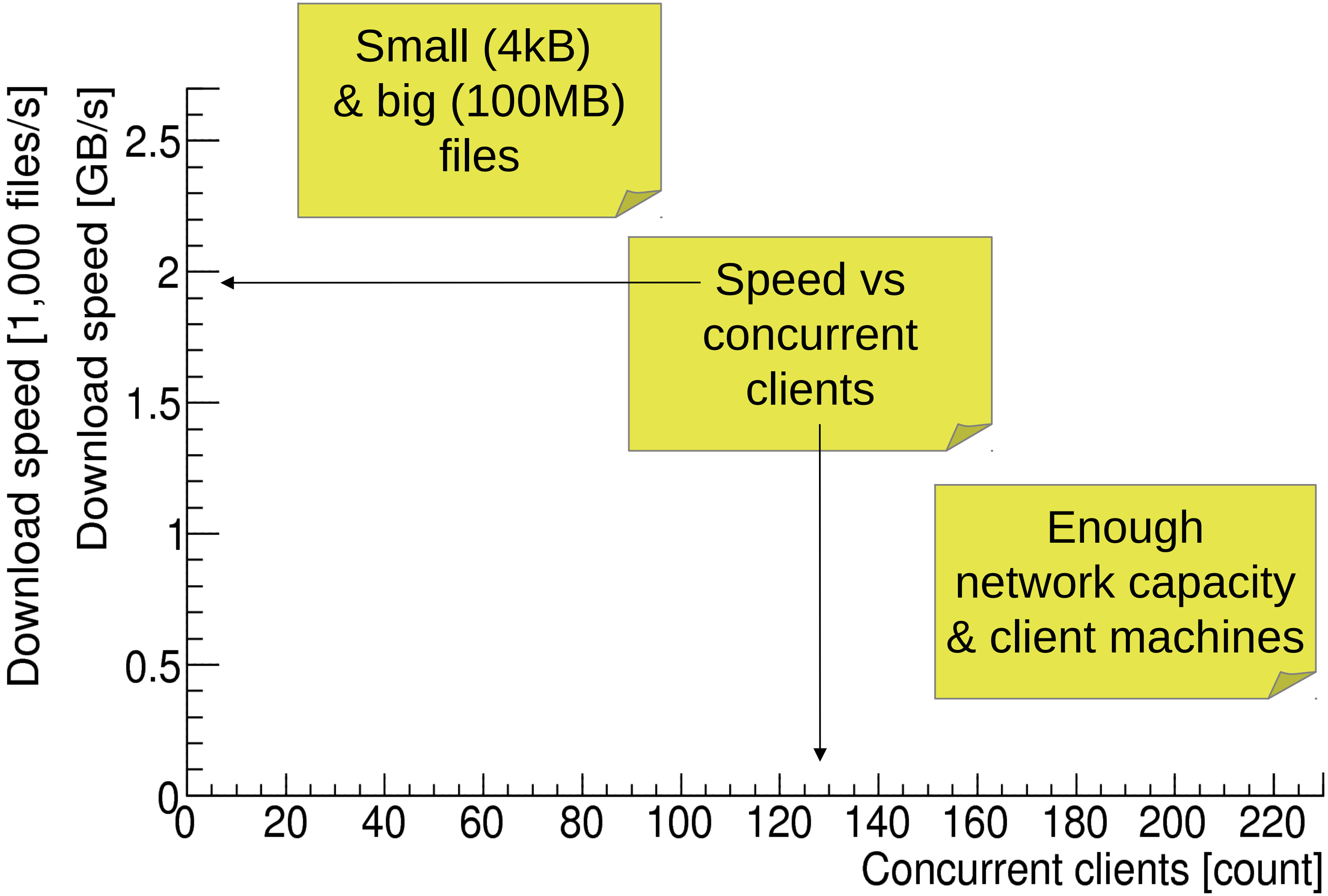
Revised version submitted in December!

Benchmarking scalability



Speed vs concurrent clients

Benchmarking scalability



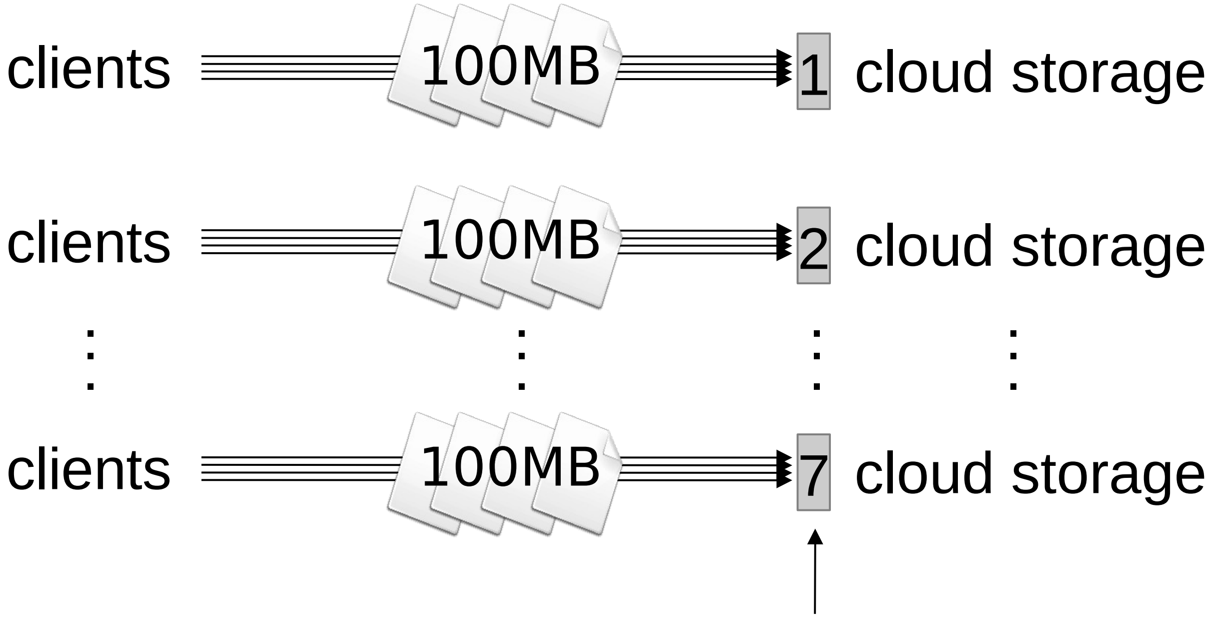
Small (4kB) & big (100MB) files

Speed vs concurrent clients

Enough network capacity & client machines

DSS

Front-end scalability



clients → 100MB → 1 cloud storage

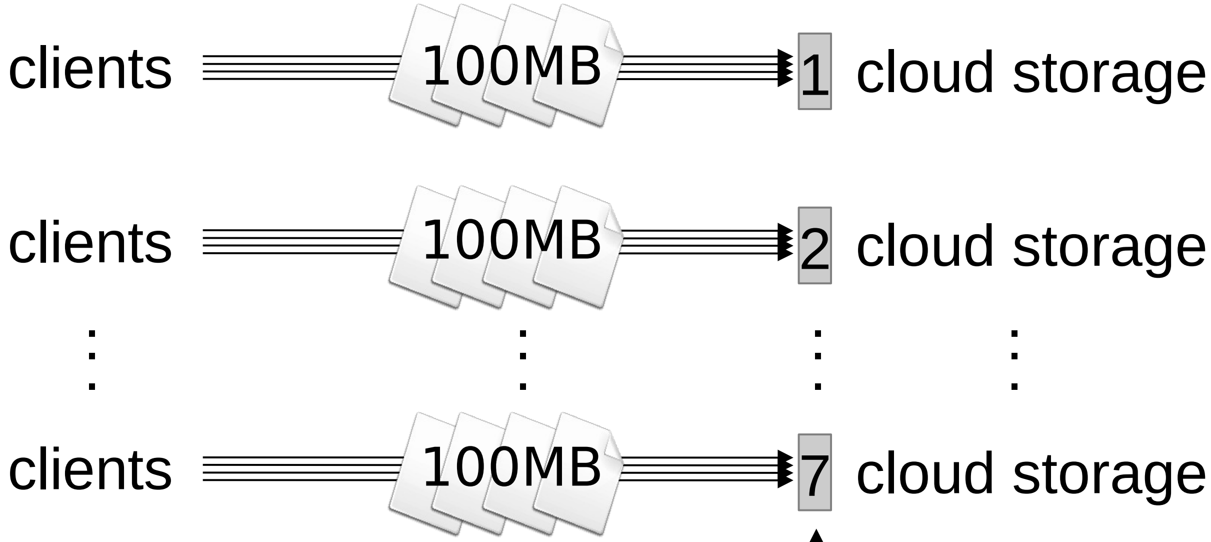
clients → 100MB → 2 cloud storage

⋮ ⋮ ⋮ ⋮

clients → 100MB → 7 cloud storage

We use different numbers of front-ends: from 1 to 7

Front-end scalability

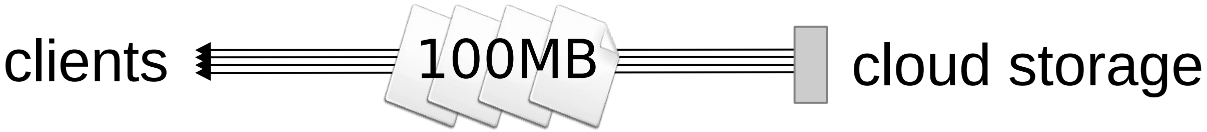


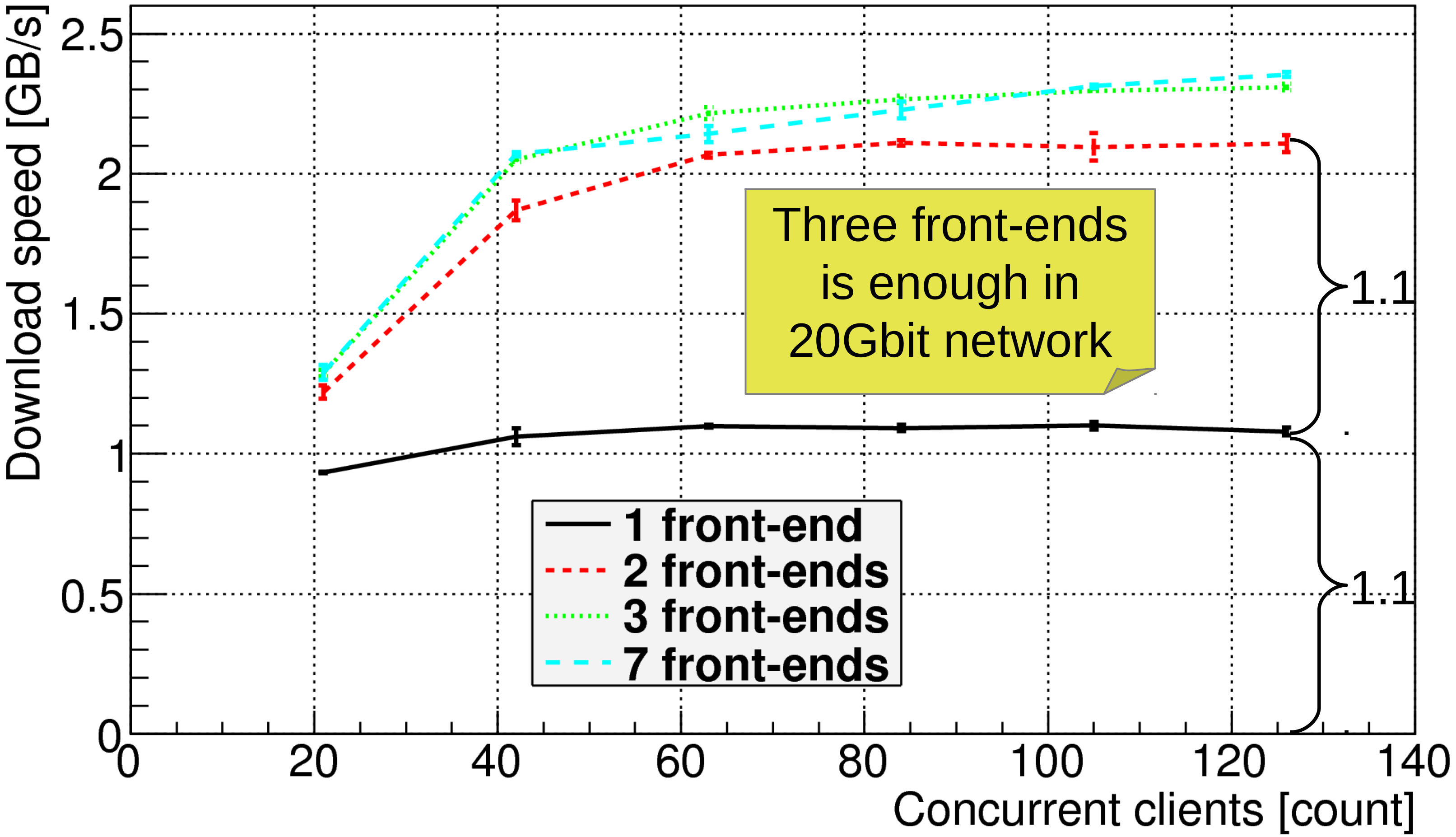
Small (4kB)
& big (100MB)
files

Uploads &
downloads

We use
different
numbers
of front-ends:
from 1 to 7

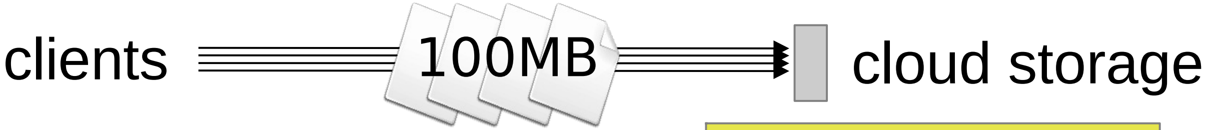
Data download scalability

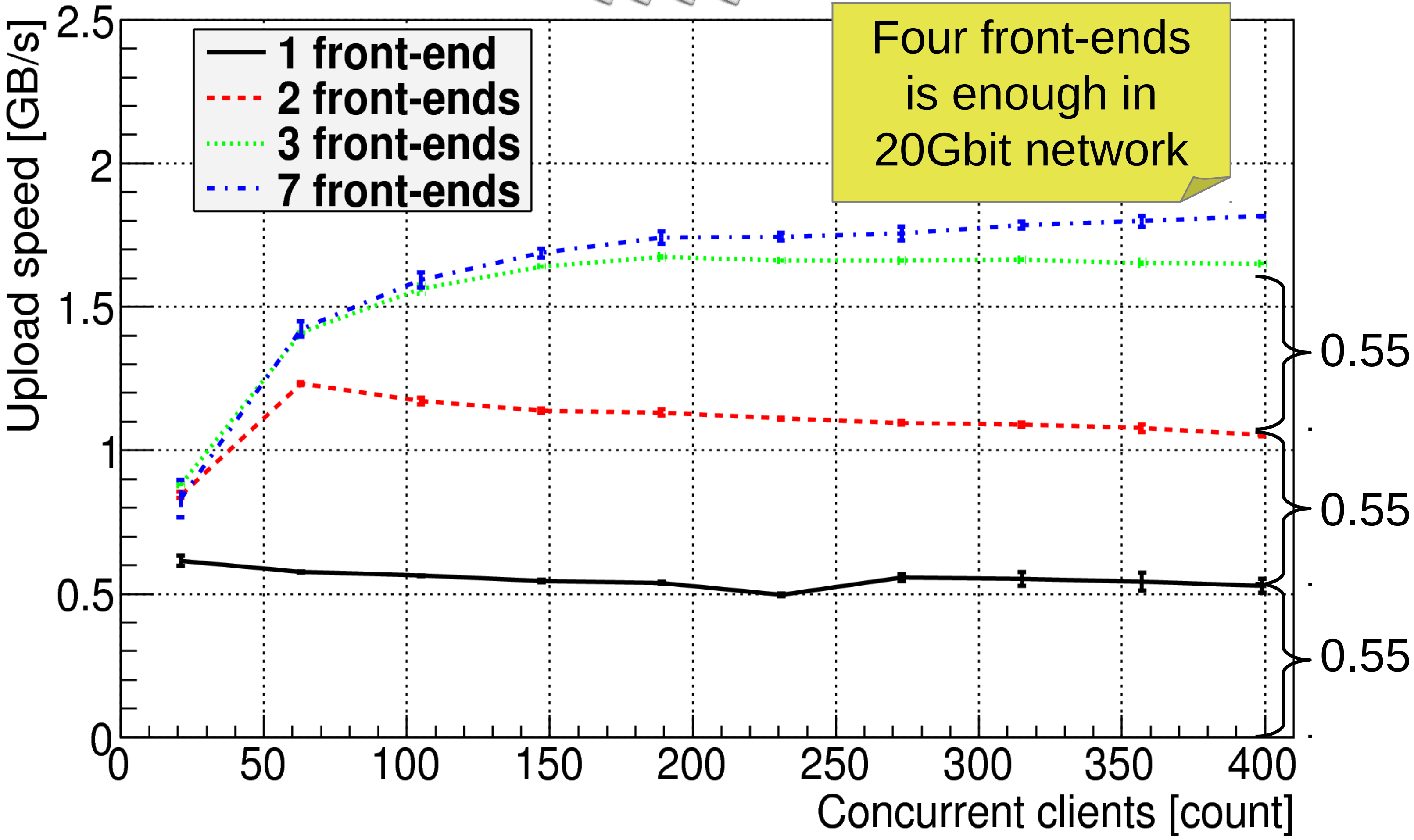




Three front-ends is enough in 20Gbit network

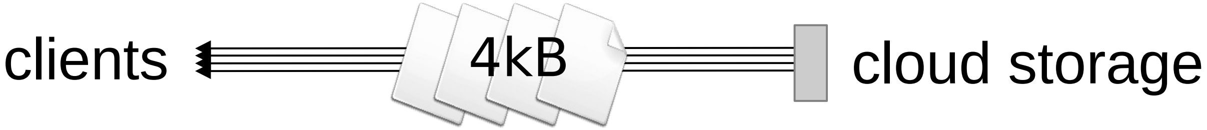
Data upload scalability

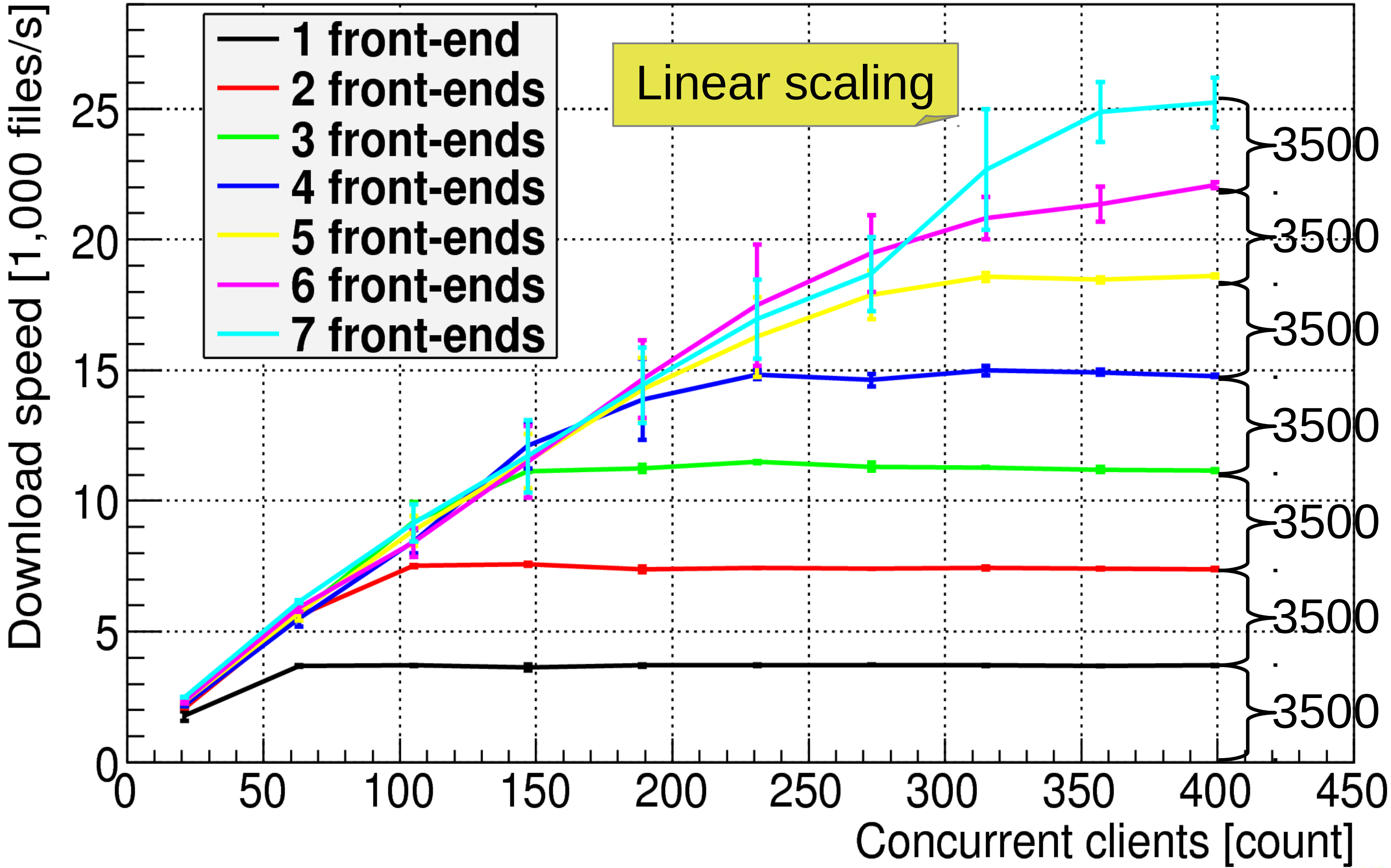




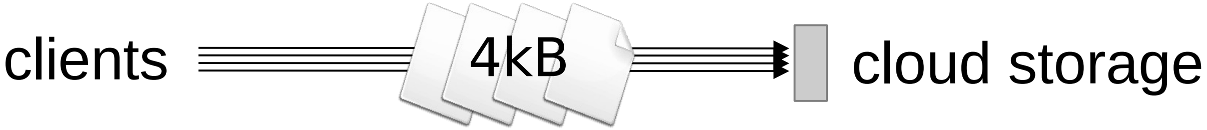
Four front-ends is enough in 20Gbit network

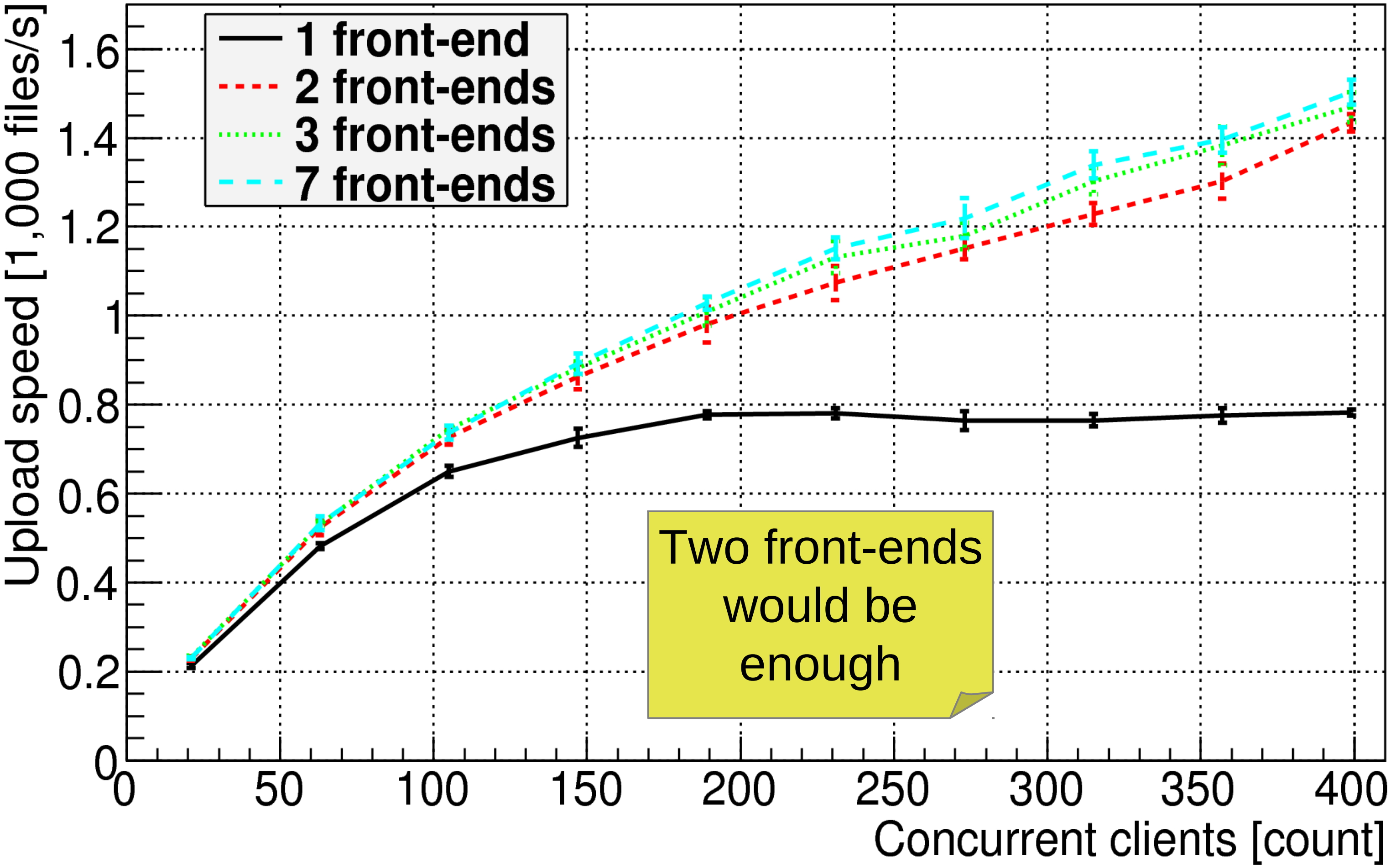
Meta-data download scalability





Meta-data upload scalability





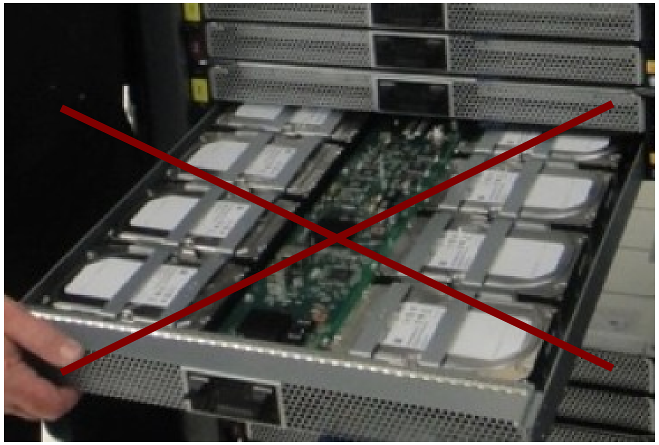
Two front-ends would be enough

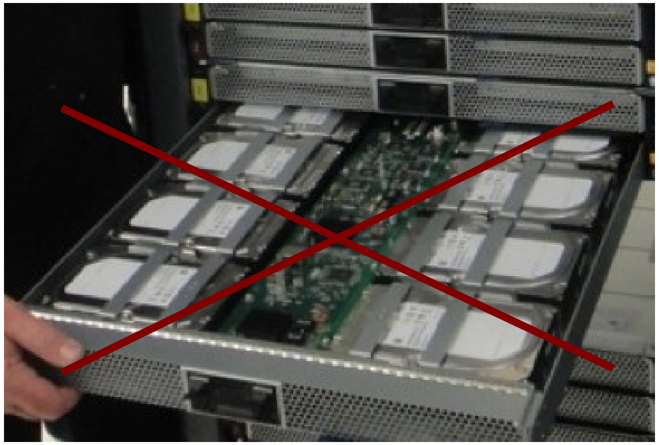
Raw performance summary

- Metadata (4kB) performance
  - 2,500 files/second upload
  - 25,000 files/second download
- Throughput (100MB) performance
  - 20Gbit network fully utilized
- Front-end scalability
  - One front-end can download 3500 files/s
  - Each front-end can upload 550 MB/s

Chassis power off recovery

Two blades are powered off:

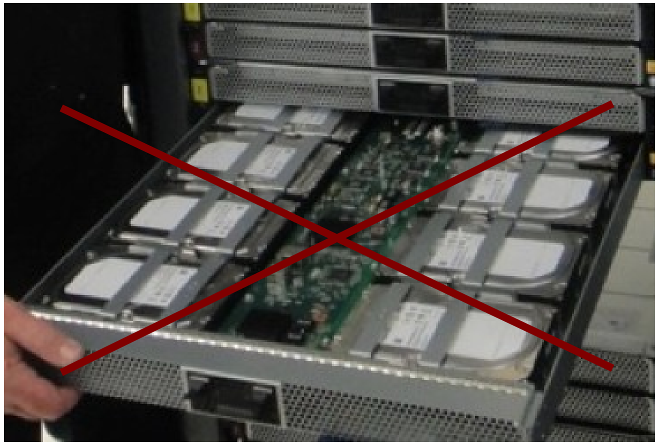


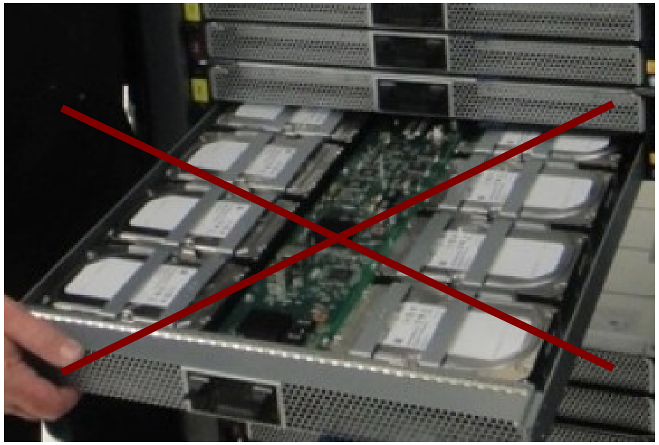


16 disks down

Chassis power off recovery

Two blades are powered off:

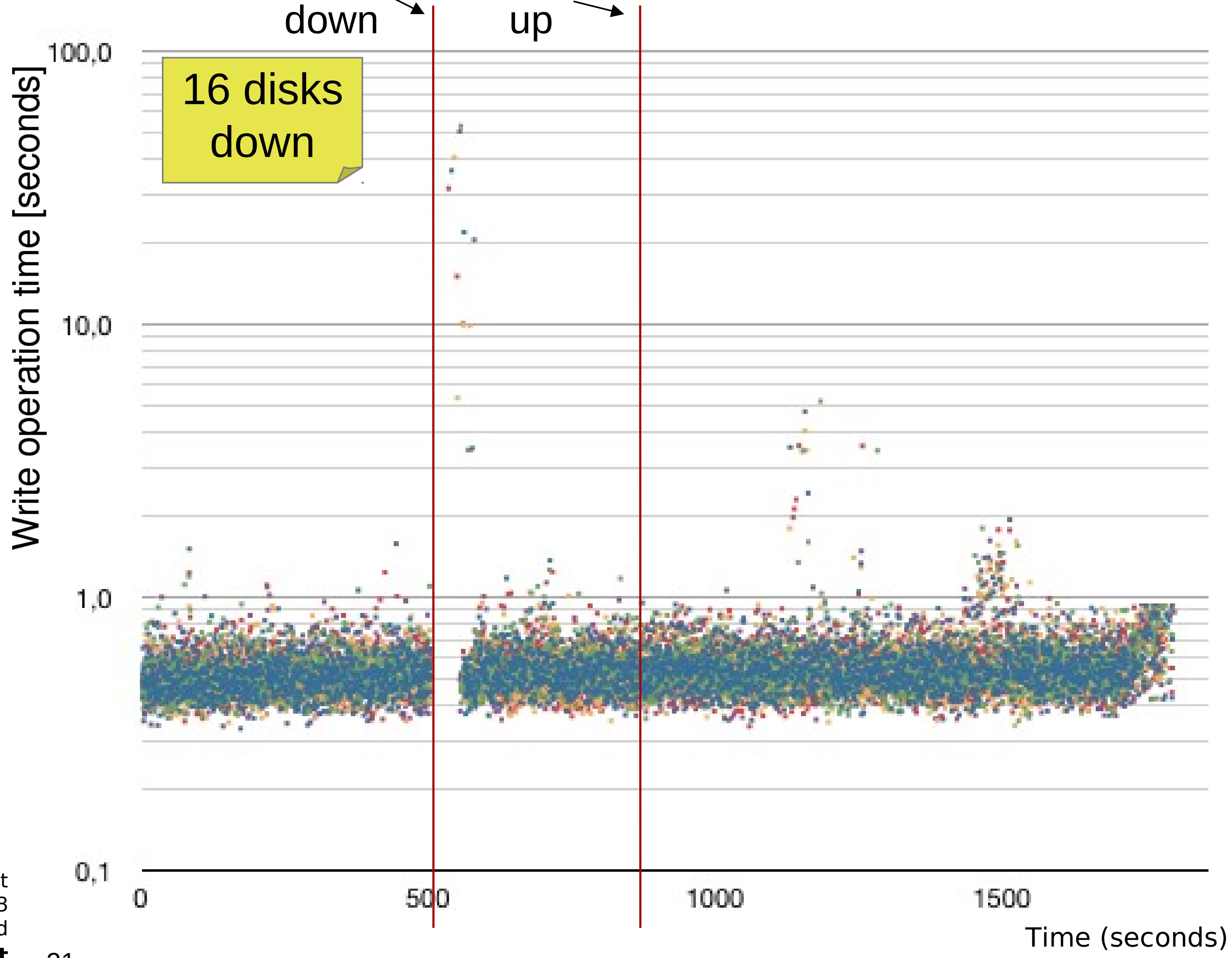




16 disks down

Uploads and downloads continue normally?

Chassis power off recovery

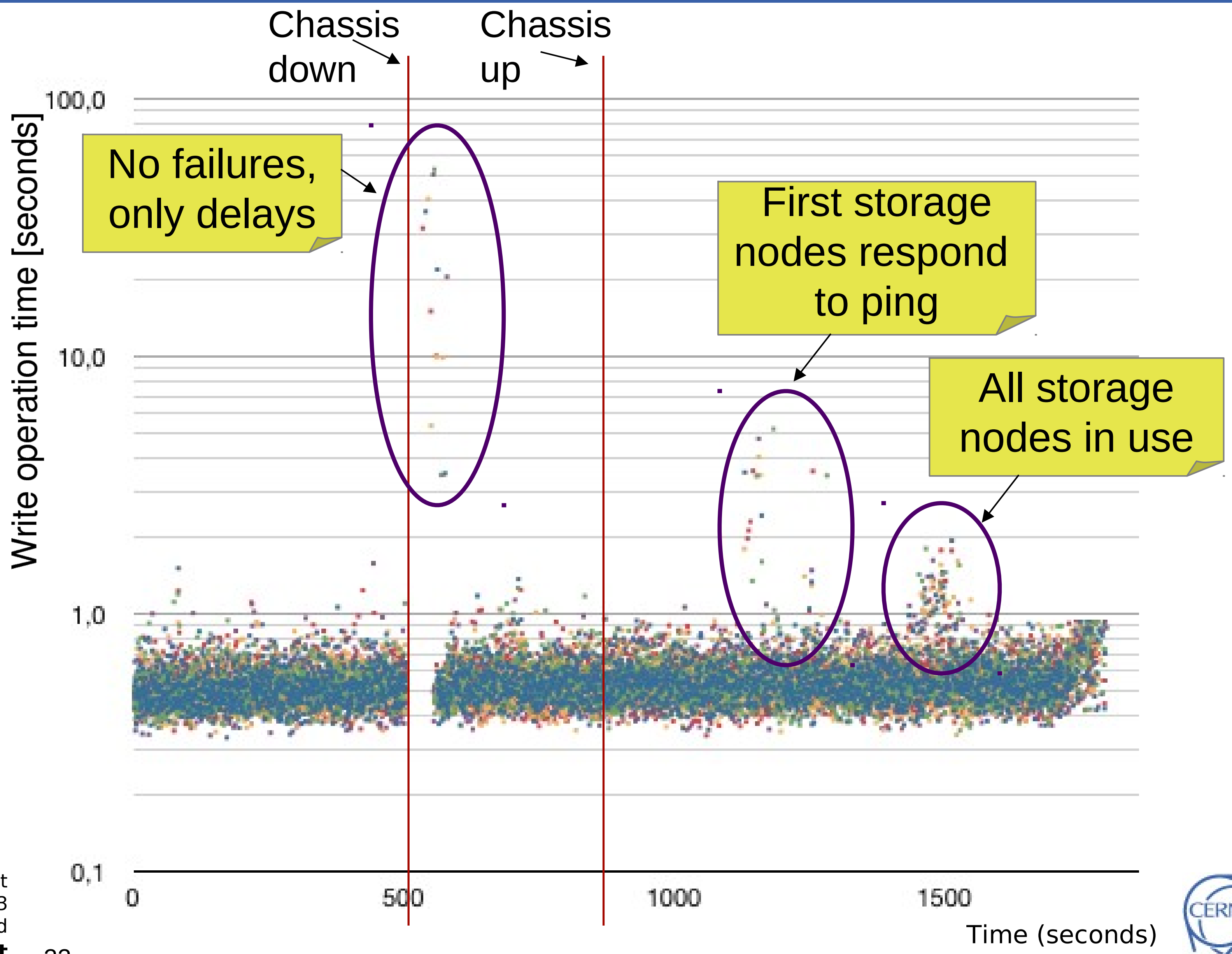


Chassis down

Chassis up

16 disks down

Chassis power off recovery



No failures, only delays

First storage nodes respond to ping

All storage nodes in use

CVMFS introduction

- What is CVMFS (CernVM File System)
  - Read only cached file system to deliver software
  - Widely used in WLCG (Worldwide LHC Computing Grid)
  - Mounted by users and files are downloaded on demand

CernVM
File system

CVMFS introduction

- What is CVMFS (CernVM File System)
  - Read only cached file system to deliver software
  - Widely used in WLCG (Worldwide LHC Computing Grid)
  - Mounted by users and files are downloaded on demand

CernVM File system

- CVMFS challenges
  - Publishing new software should be fast (upload tens of thousands of files)
  - Files should be accessed with HTTP protocol

File system back-end

CernVM File system

- Implementation
  - Files are uploaded to multiple buckets in the cloud storage
  - Files are downloaded with unified name space

    http://cloud.cern.ch/~~bucket-42/~~file001.bin

    http://cloud.cern.ch/file001.bin

File system back-end

CernVM File system

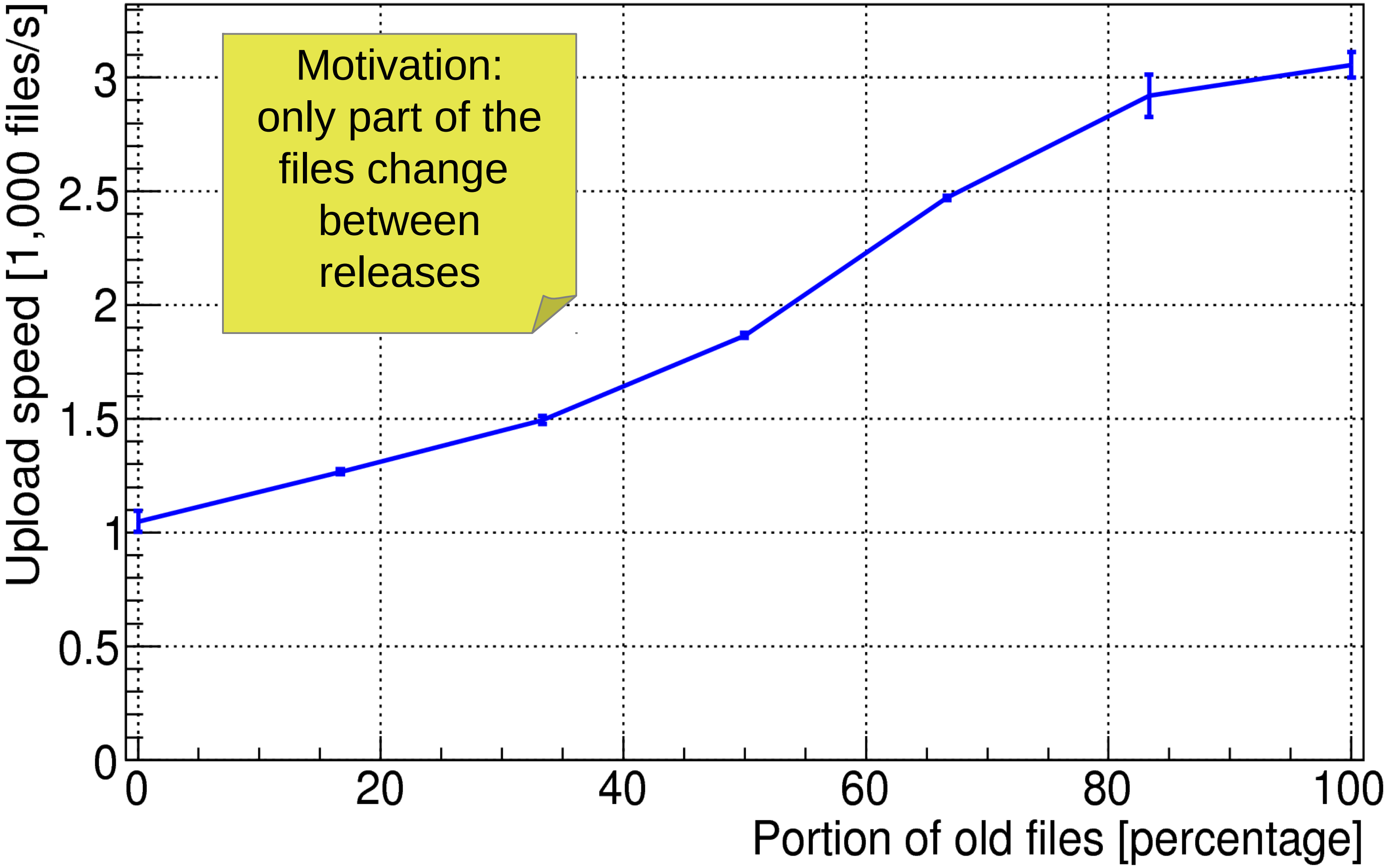
- Implementation
  - Files are uploaded to multiple buckets in the cloud storage
  - Files are downloaded with unified name space

    http://cloud.cern.ch/~~bucket-42/~~file001.bin

    http://cloud.cern.ch/file001.bin
- Result
  - Full publish procedure tested to work using 30,000 small files
  - Upload speed 1200 files/second (with 240 threads)

CVMFS incremental uploads

Uploading 30,000 files (of average size 10kB)
to Huawei cloud storage



Motivation:
only part of the
files change
between
releases

CHEP paper summary

- Raw performance
  - Upload and download scalability demonstrated
  - Additional front-end nodes increased linearly the performance
- Fault tolerance: powering off a chassis
  - Transparent disk failure recovery demonstrated
- File system with cloud storage back-end
  - Full publishing procedure tested
  - Uploading of only new files feature tested

Multi-part uploads

- Problem: no multi-part upload support ?
  - CVMFS software is foreseen to require multi-part uploads to S3 cloud storages in the near future
- Solution: supported in the new version !
  - Current version of the Huawei cloud storage in CERN does not support multi-part uploads, but latest version does
  - New version will be tested when deployed in CERN

New ROOT plugin tests

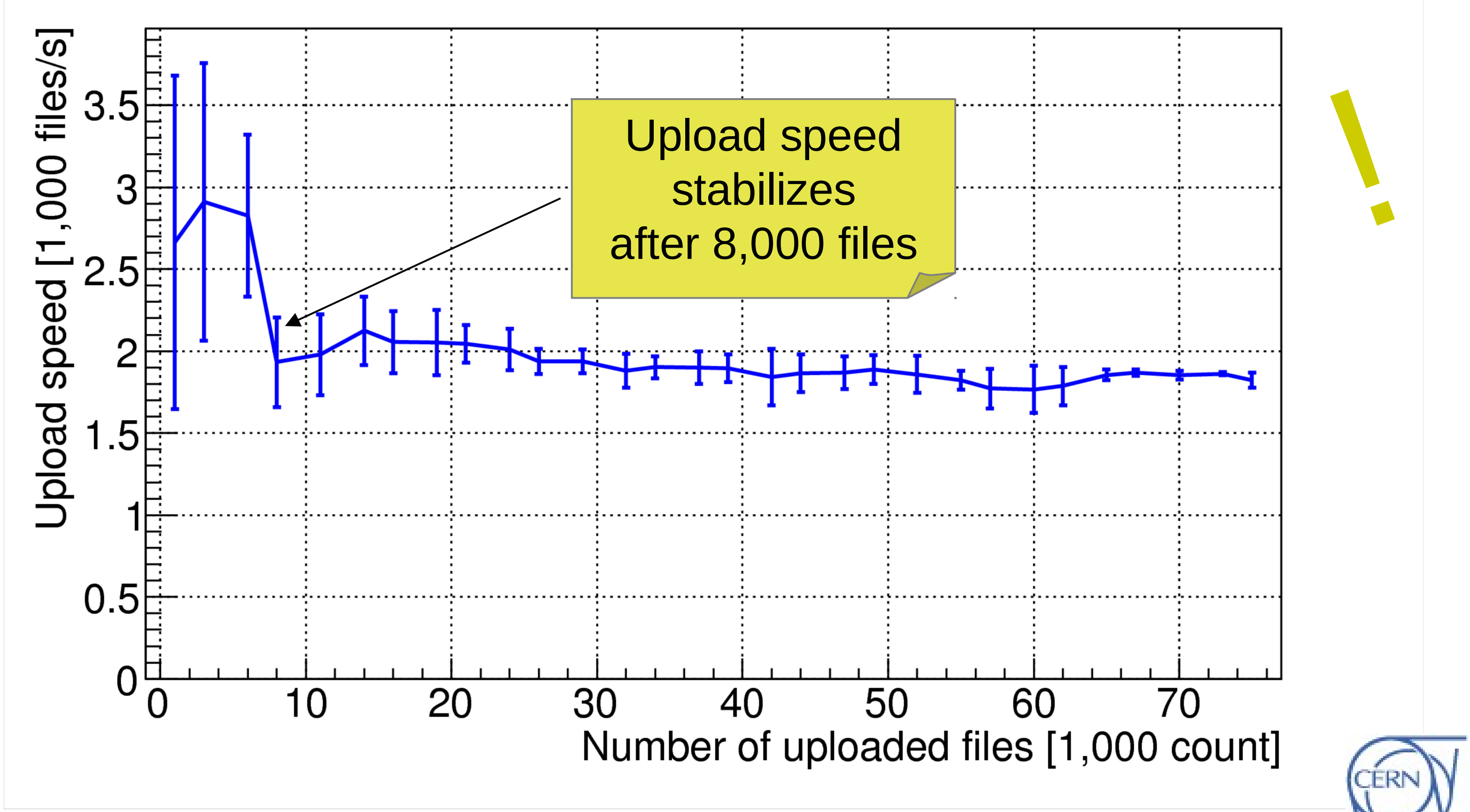
- Problem: new ROOT S3 plugin ?
  - – New ROOT S3 plugin was released, is it working properly with Huawei cloud storage?

- Solution: tested to work with one client !
  - – One client read-performance identical to the old ROOT S3 plugin
  - – Multi-client stress tests are planned

Test: Consecutive uploads

- Problem: consecutive uploads
  - Does number of consecutive uploads affect the upload speed

?



Upload speed stabilizes after 8,000 files

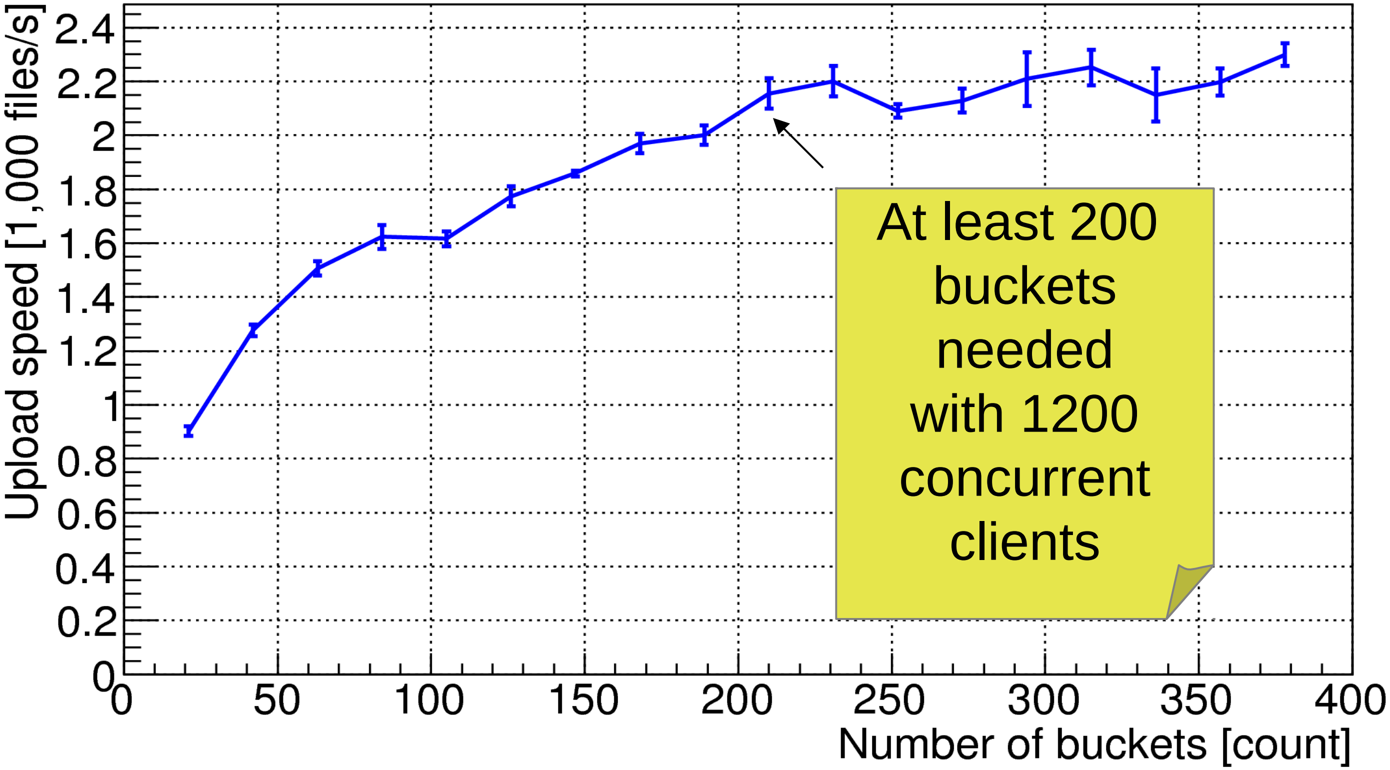
!

Test: Number of buckets

- Problem: how many buckets needed
  - How the number of used buckets affects the maximum achievable upload speed

?

!

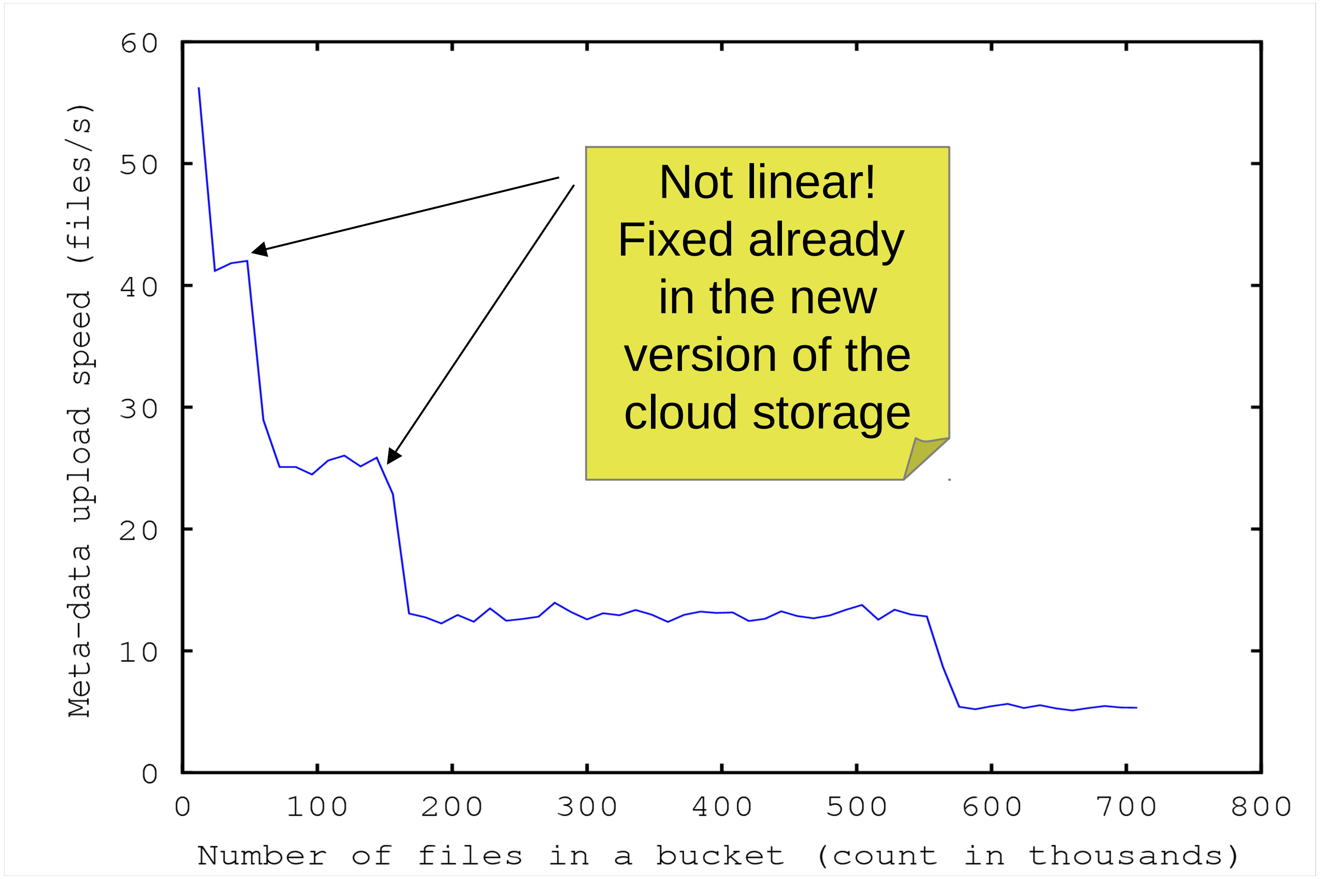


At least 200 buckets needed with 1200 concurrent clients

Test: Many files in one bucket

- Problem: bucket usage affects speed?
  - How the number of files in a bucket affects the maximum achievable upload speed

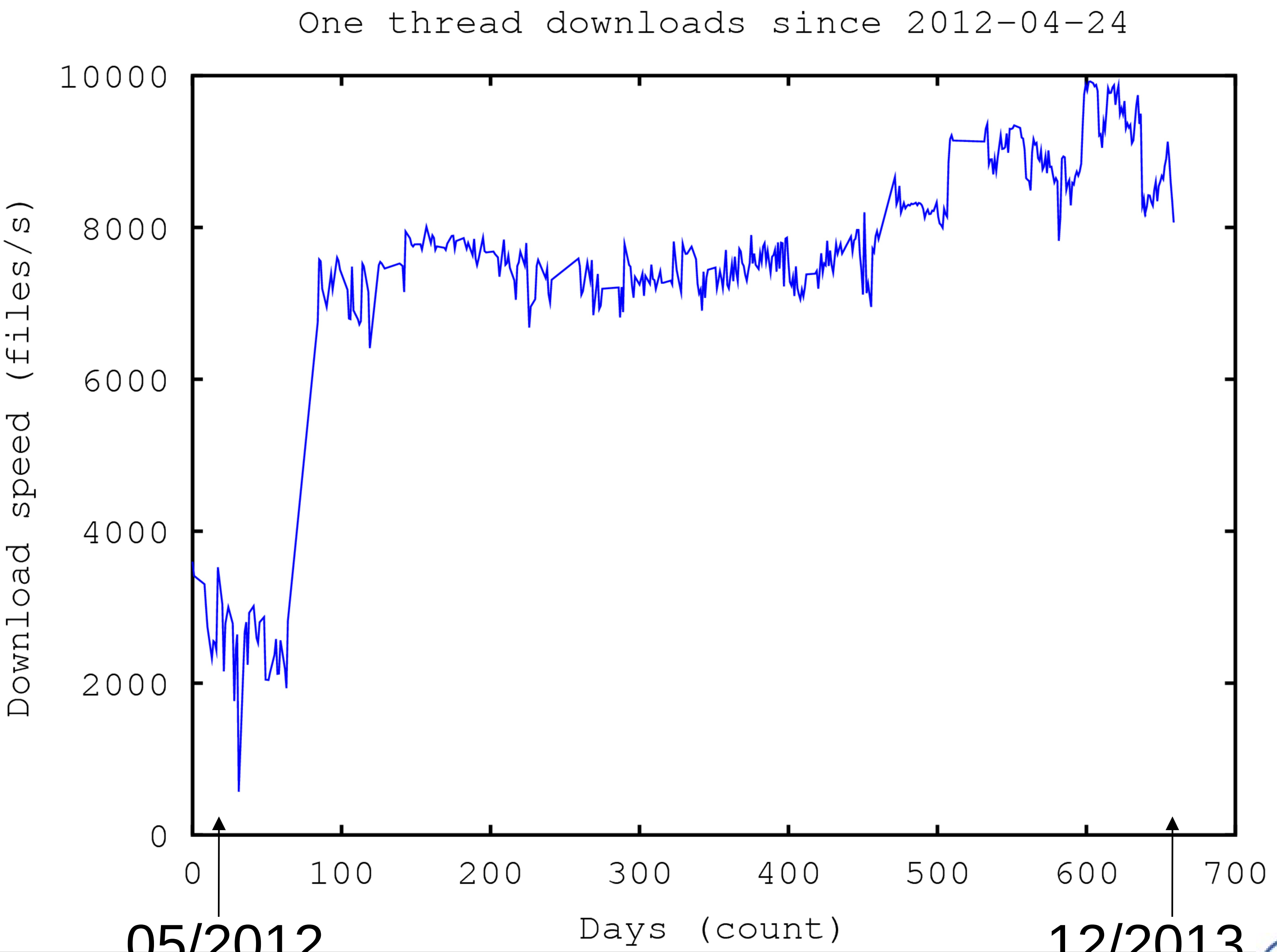
?



Not linear! Fixed already in the new version of the cloud storage

!

Long term download stability



05/2012

12/2013

Conclusions

- Identified new requirements
  - Multi-part file upload support
  - Bucket fullness should not affect the upload performance
  - Fast upload speeds should not require hundreds of buckets

Fixed already in the new version!

- Test results
  - New ROOT S3 plugin worked without problems with the Huawei cloud storage
  - Long-term download stability good

Future plans

- Short term
  - Benchmark CVMFS with real release data
  - Test ROOT S3 plugin performance with multiple clients
- Long term
  - Second petabyte system with enterprise disks expected to arrive soon
  - Replication tests between cloud storages
  - Prove total cost of ownership (TCO) gains of the system as part of a production service

Future plans

- Short term
  - Benchmark CVMFS with real release data
  - Test ROOT S3 plugin performance with multiple clients
- Long term
  - Second petabyte system with enterprise disks expected to arrive soon
  - Replication tests between cloud storages
  - Prove total cost of ownership (TCO) gains of the system as part of a production service

Thank you!

seppo.heikkila@cern.ch

Huawei Cloud Storage

Seppo S. Heikkila
Maria Arsuaga Rios
CERN IT

Openlab Major Review Meeting
13th of February 2014
CERN, Geneva

HUAWEI

CERN openlab