DSS

Data & Storage Services

CERN IT Department

Huawei Cloud Storage Evaluation and Testing with Prototype Services

Seppo S. Heikkila
CERN IT

Openlab Minor Review
29th of October 2013
CERN, Geneva

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

HUAWEI

CERN

CERN openlab

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

Huawei cloud storage

CERN IT Department

DSS

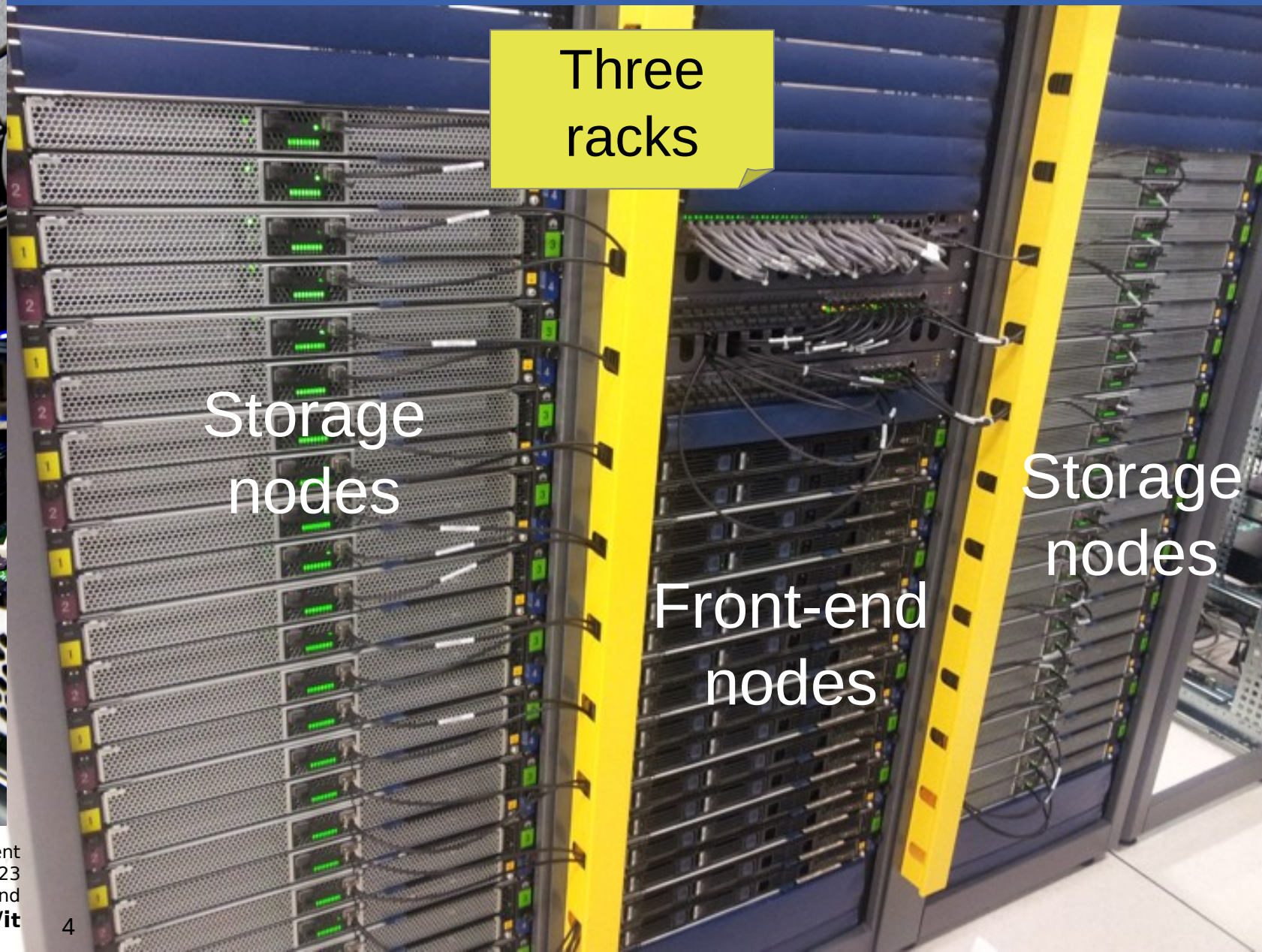
Location: CERN Computer Center

"We are now here"

"Cloud storage"

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

Huawei cloud storage setup



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

Huawei cloud storage setup

384 disks → 768 TB

Storage nodes

S3 compatible

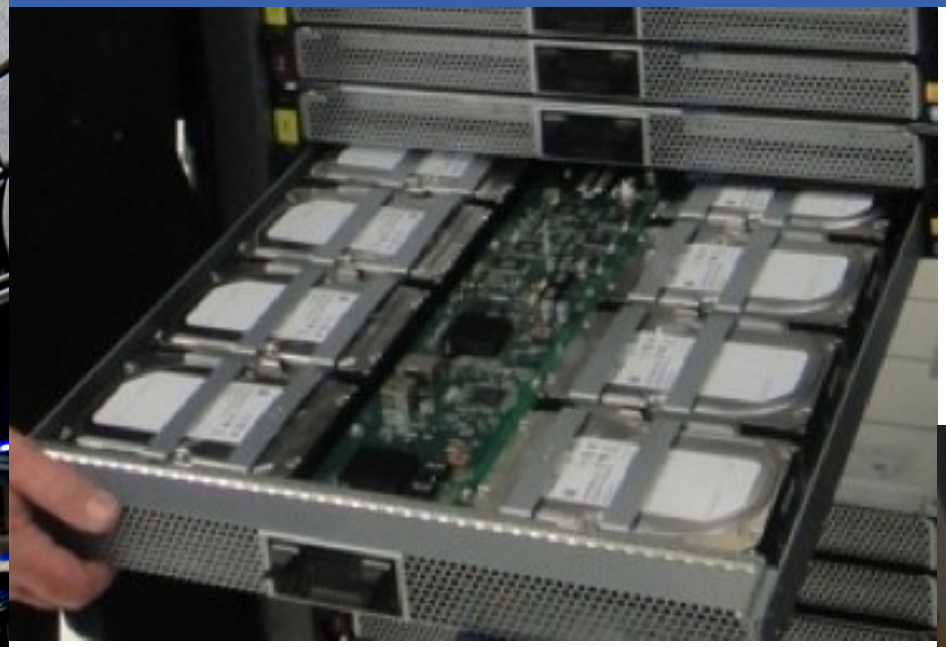
Buckets divide the name space

Front-end nodes

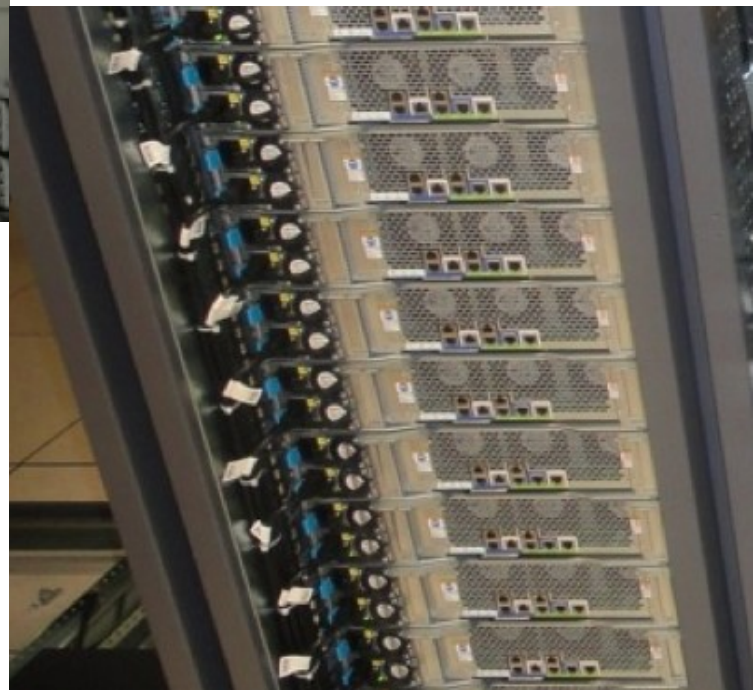
Storage nodes

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

Storage nodes



Each blade has eight storage nodes



One chassis has two blades (16 disks)

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

Timeline

1.5 years of Huawei...

Major Review

Major Review

Major Review

Major Review

Minor Review

Board of Sponsors

Minor Review

Board of Sponsors

Minor Review

01/2012

01/2013

10/2013

Project starts

First tests

Upgrade of the system

Stress testing

File-system integration

Commissioning of the system

Failure recovery testing

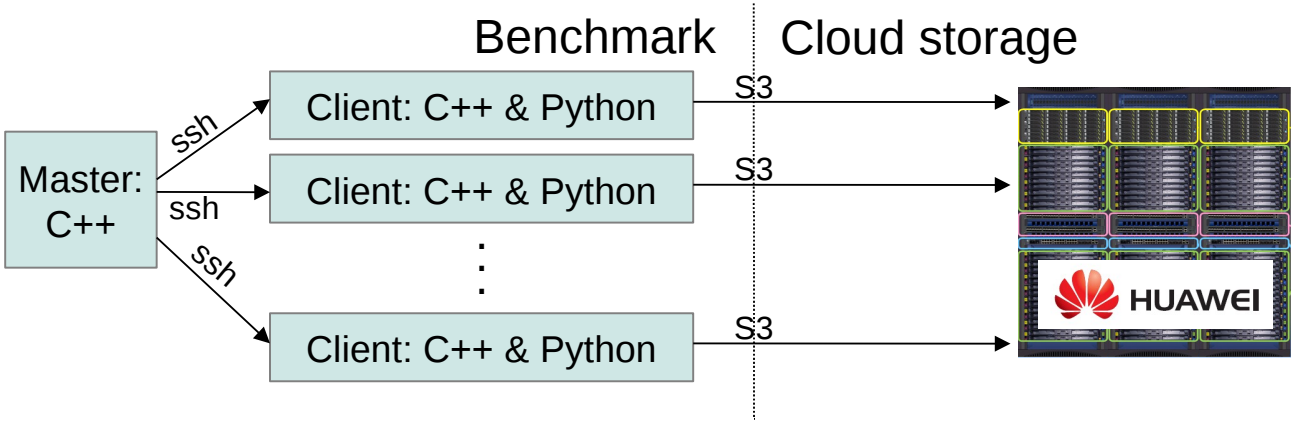
Full-scale stress testing

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

Benchmark setup

Distributed C++ benchmark

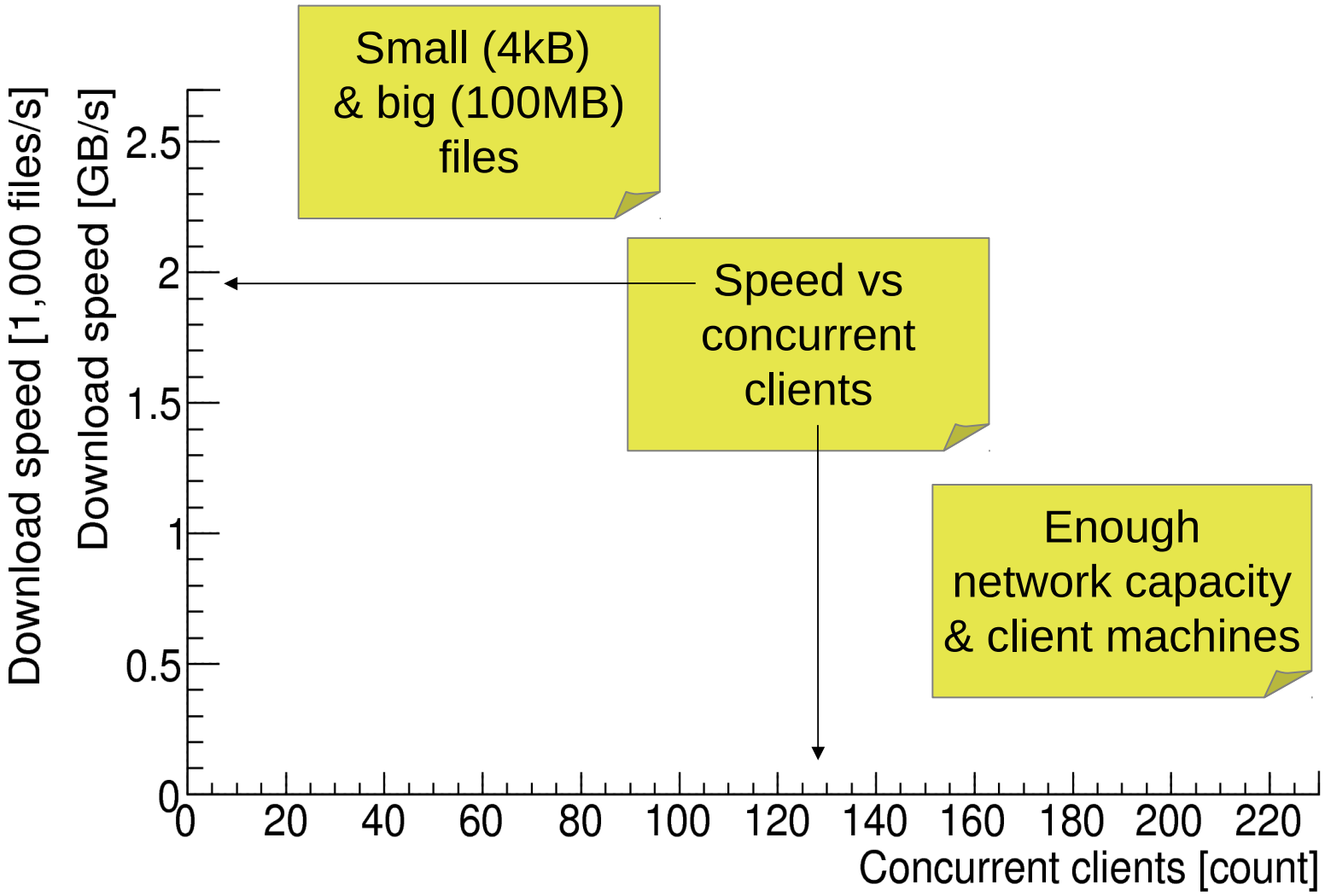
- Integrated with ROOT
- Client nodes connected with ssh
- S3 Python library to read and write files
- Histograms about specific metrics
  - Operation time, read/write speed, CPU/memory utilisation



Benchmarking scalability

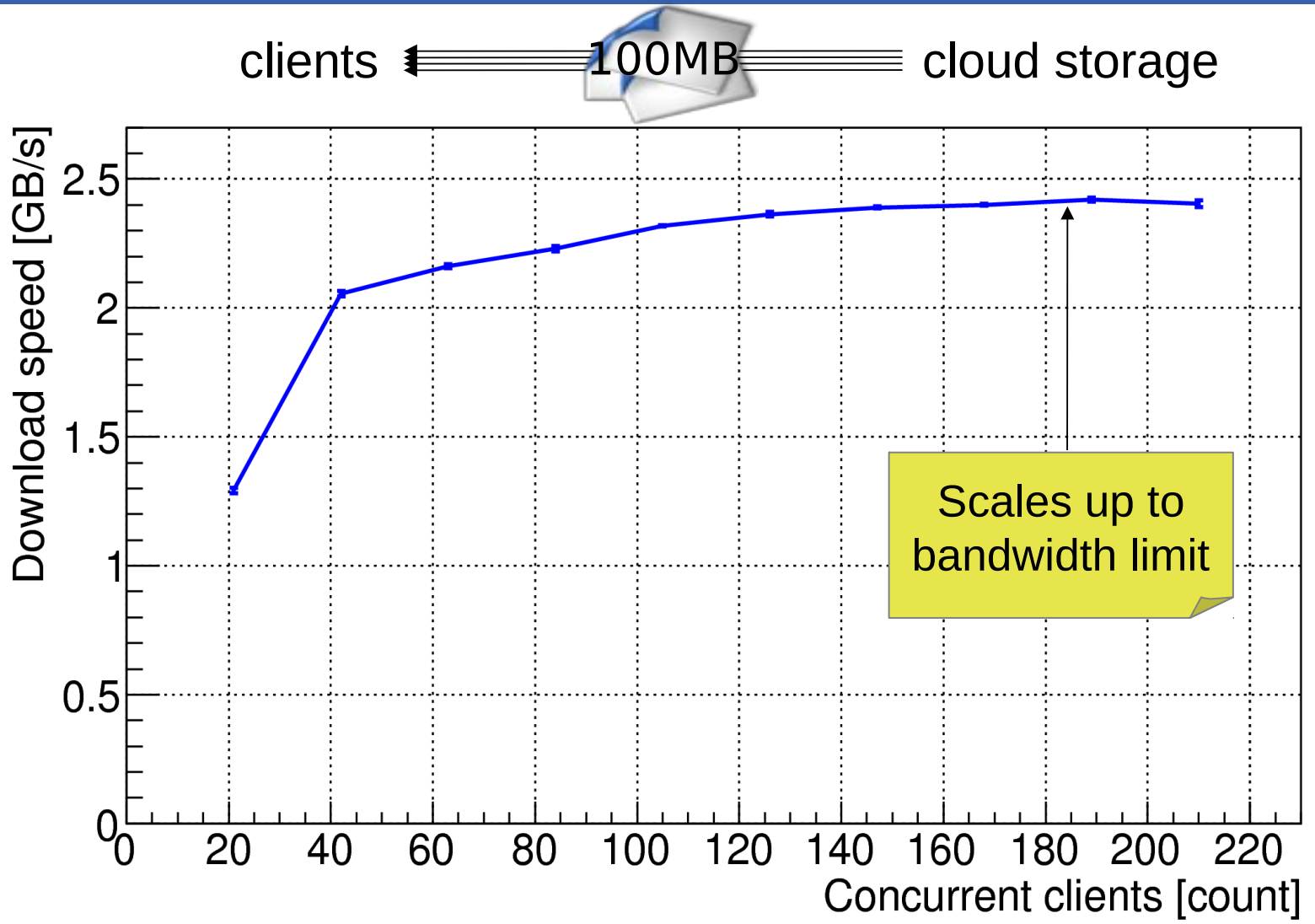
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Benchmarking scalability



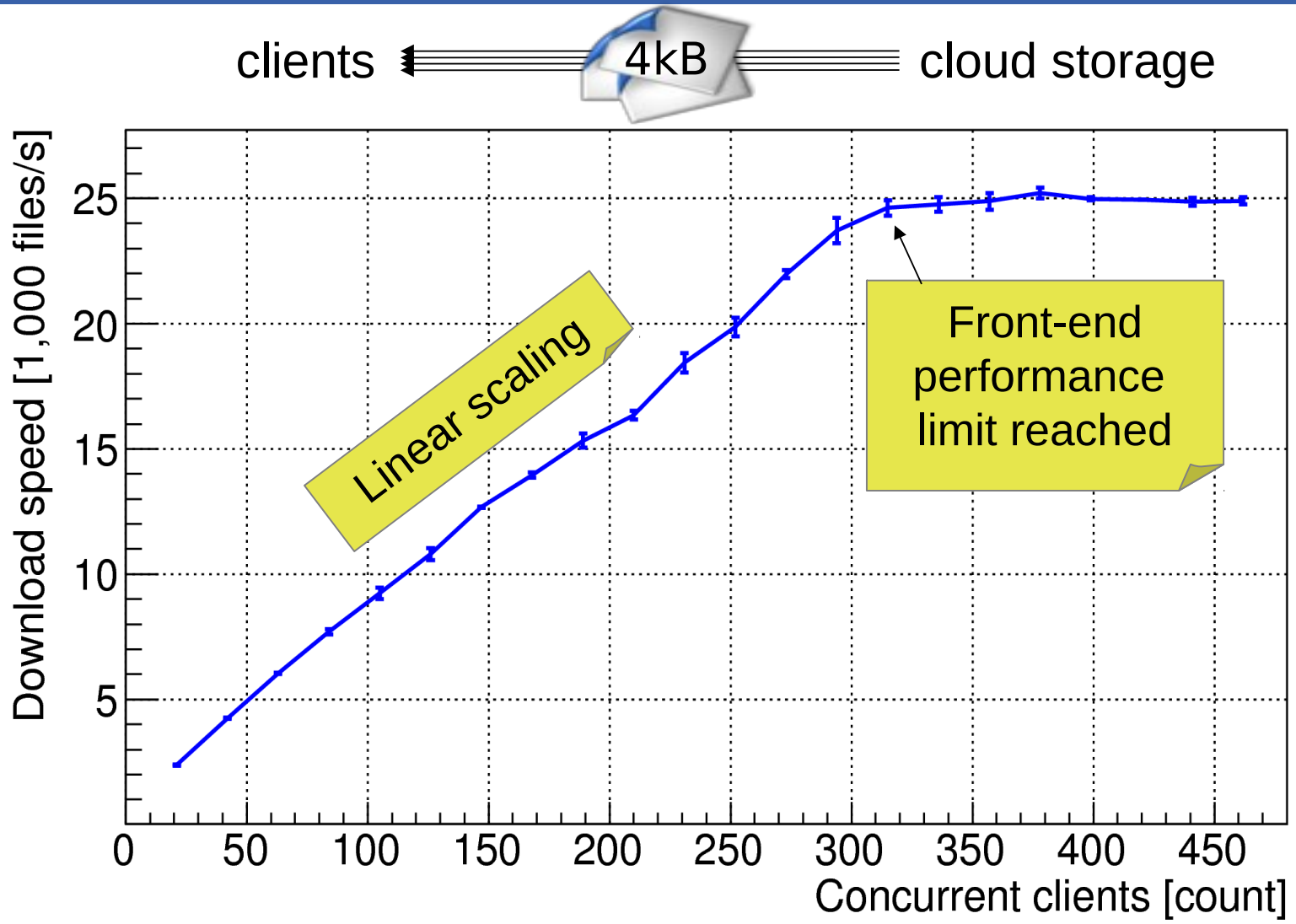
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Download throughput scalability



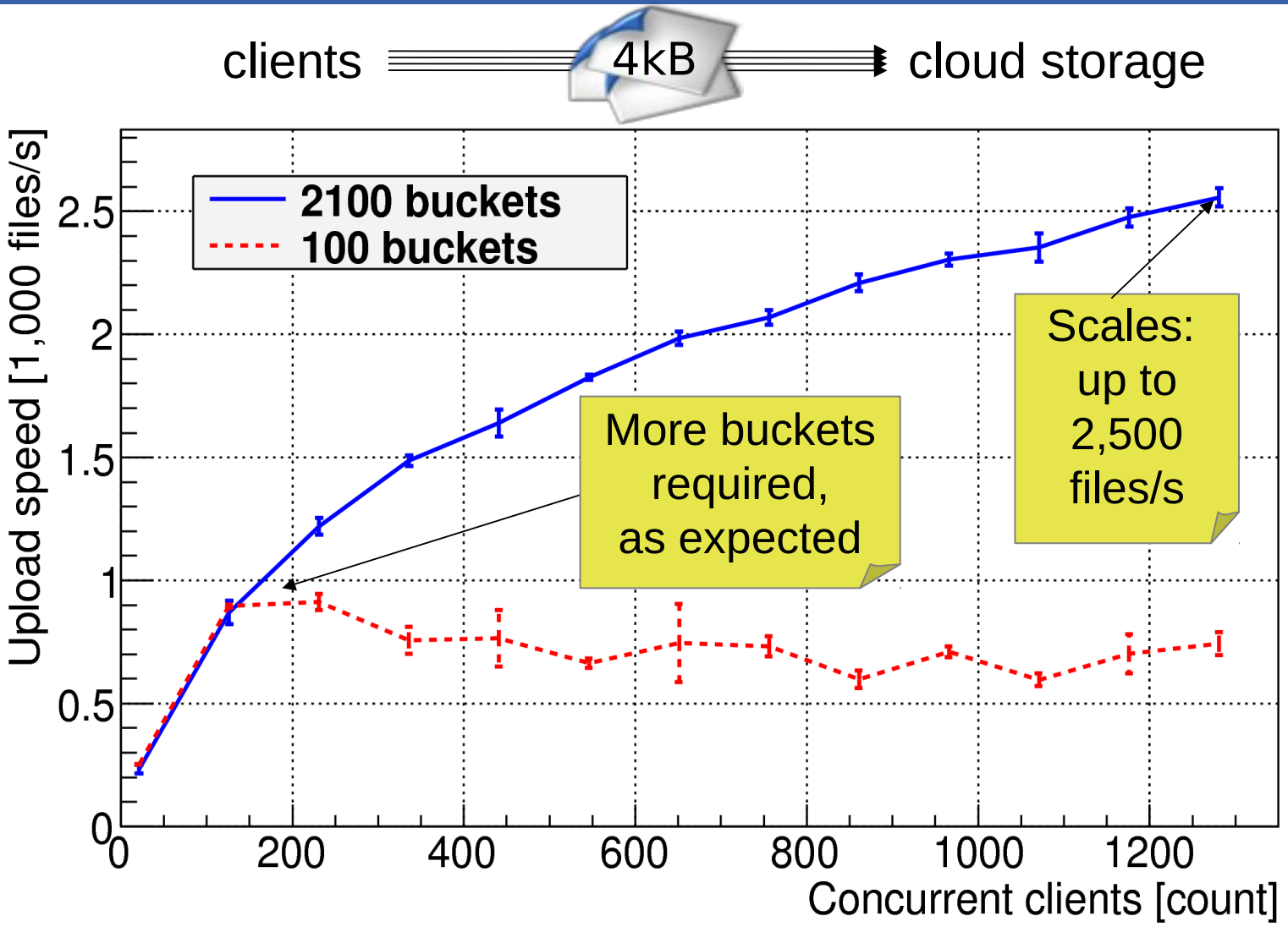
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Download metadata scalability



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

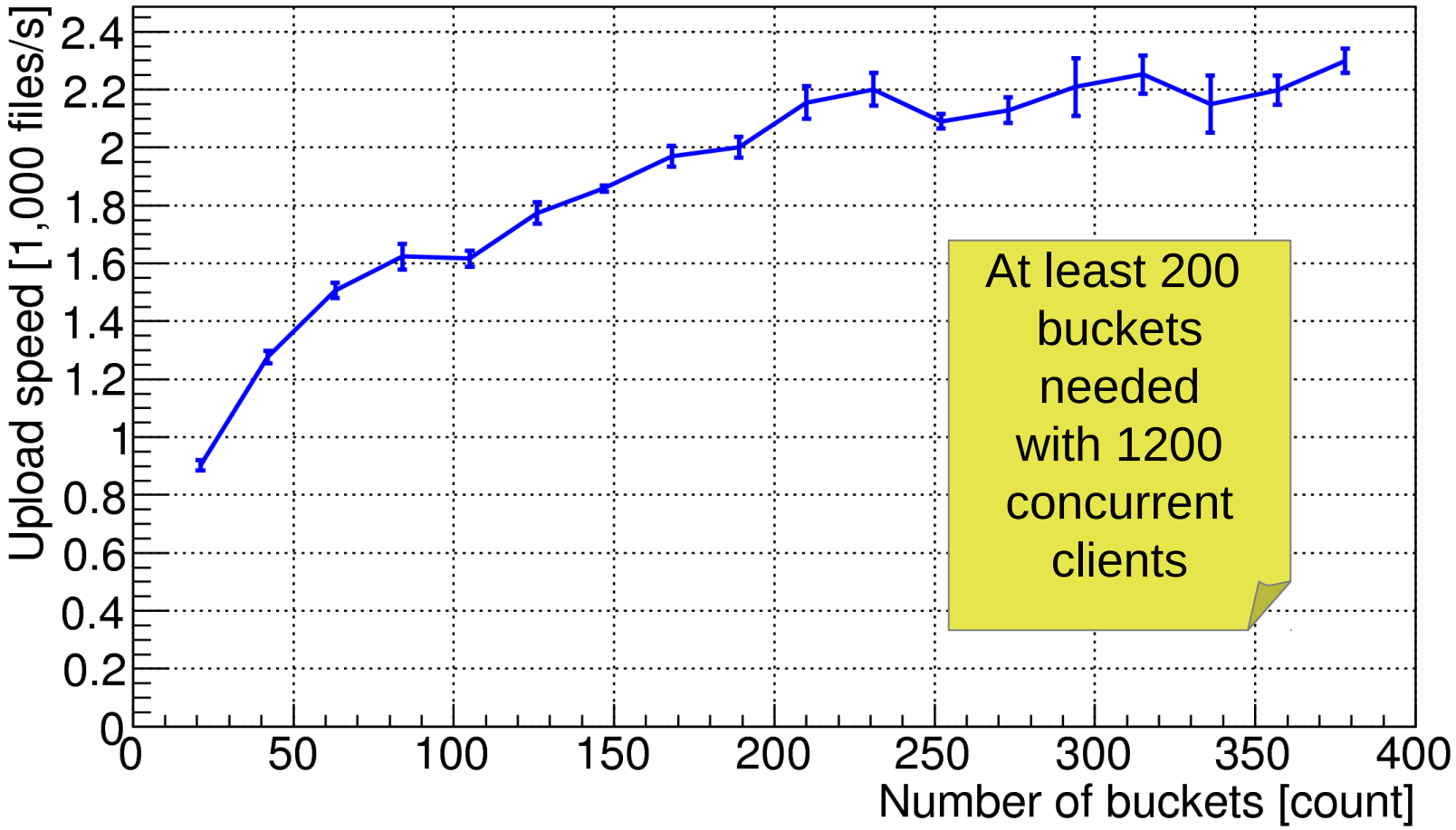
Upload metadata scalability



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

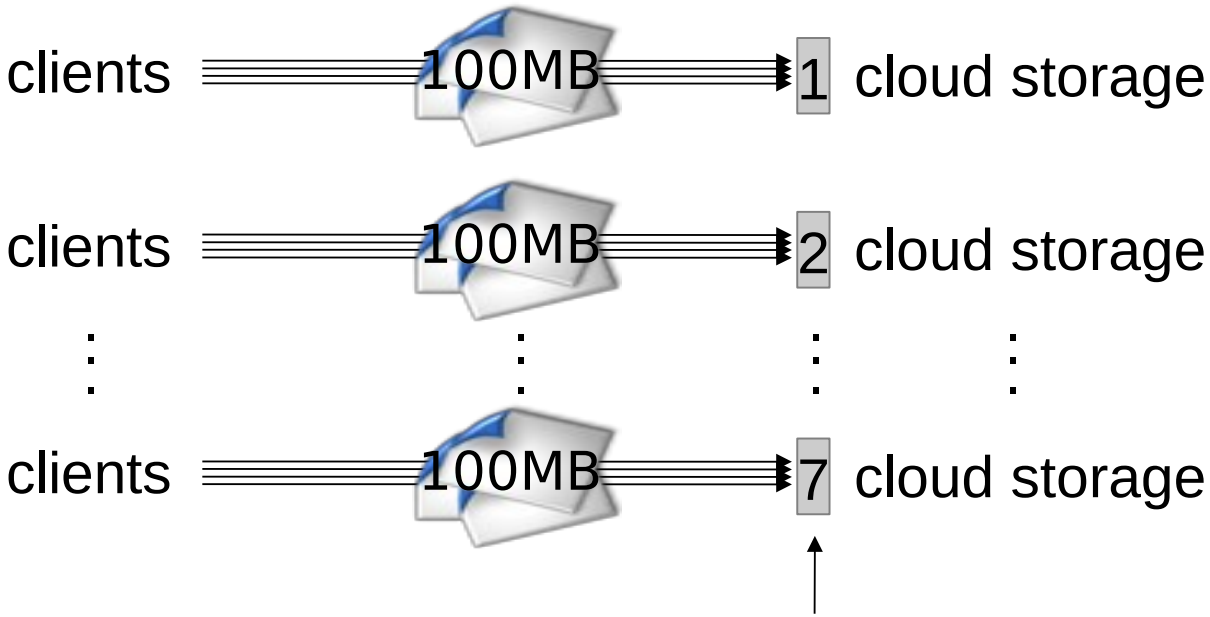
DSS Number of buckets vs upload speed

CERN IT Department



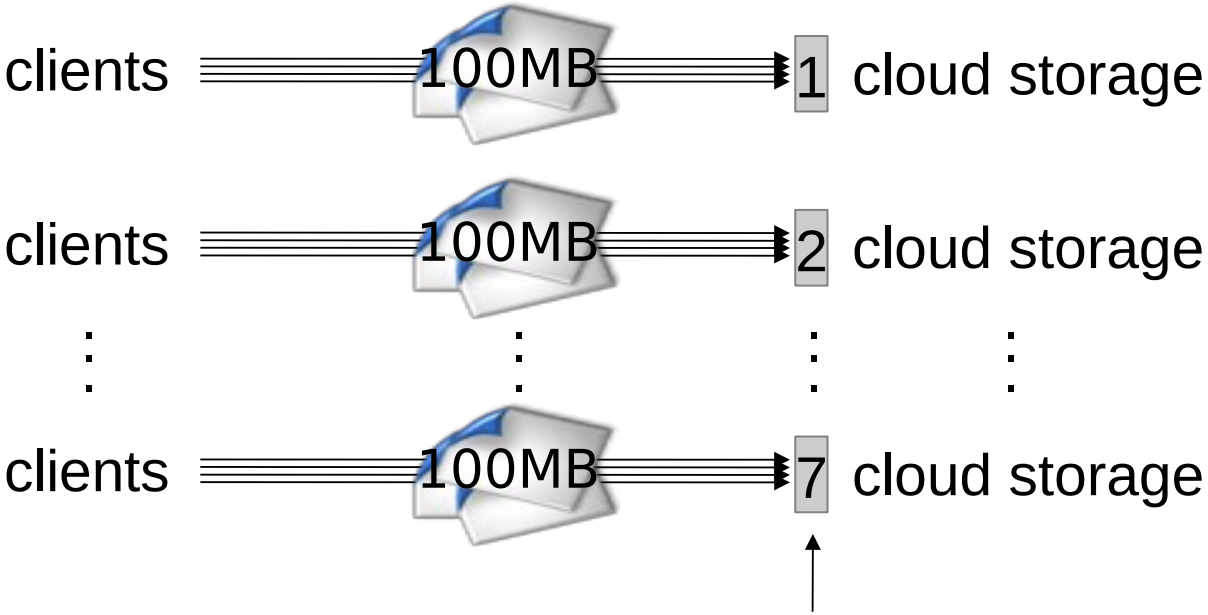
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Front-end scalability



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

Front-end scalability



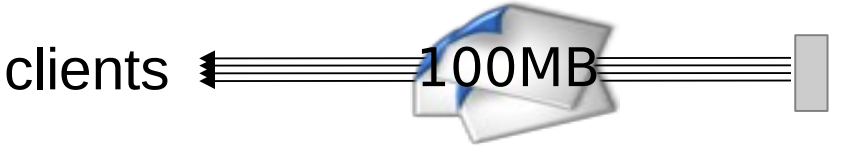
Small (4kB) & big (100MB) files

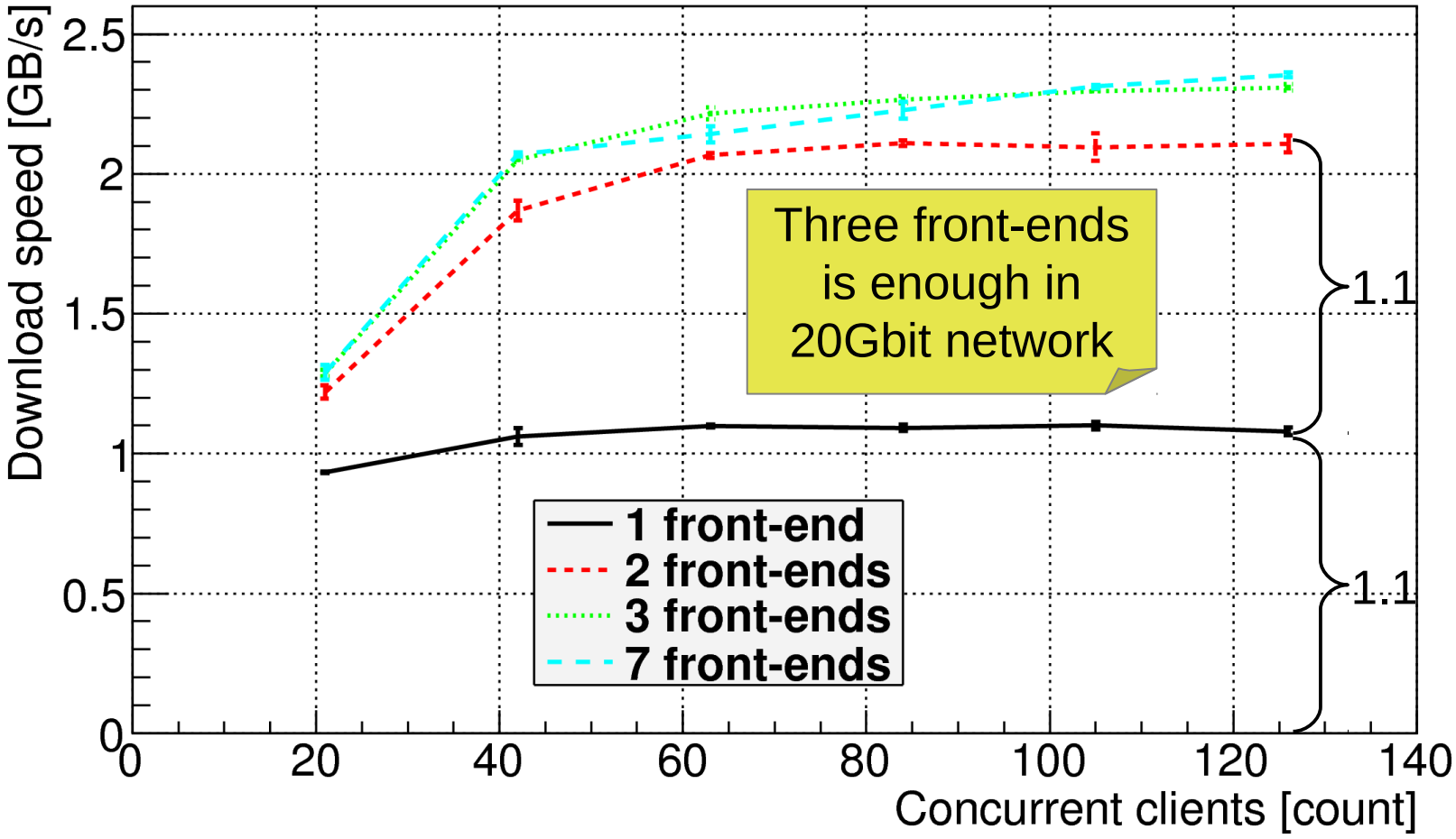
Uploads & downloads

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

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

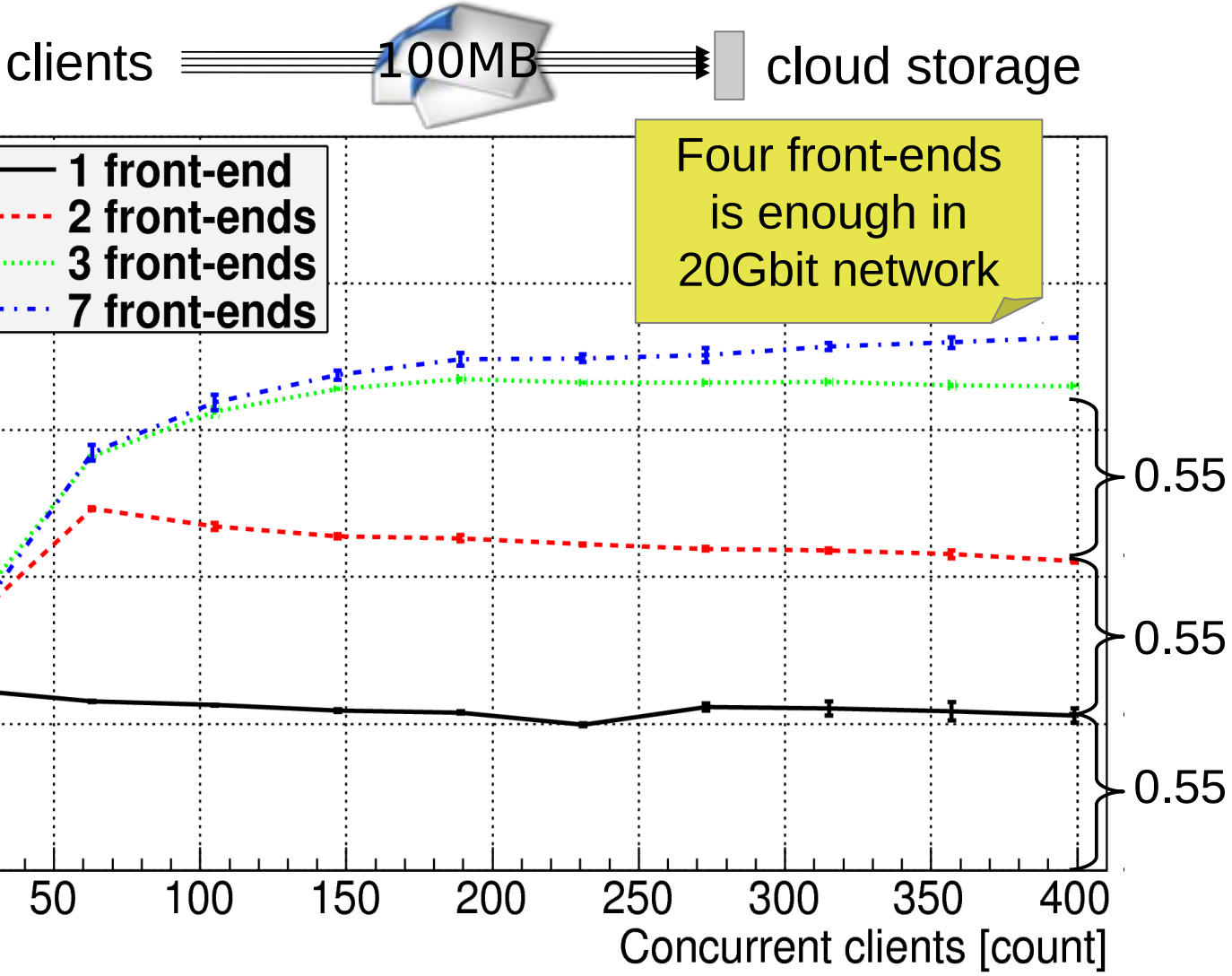
Front-end scalability





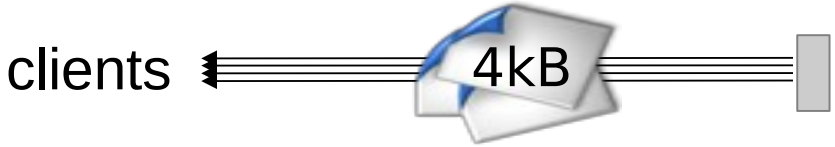
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

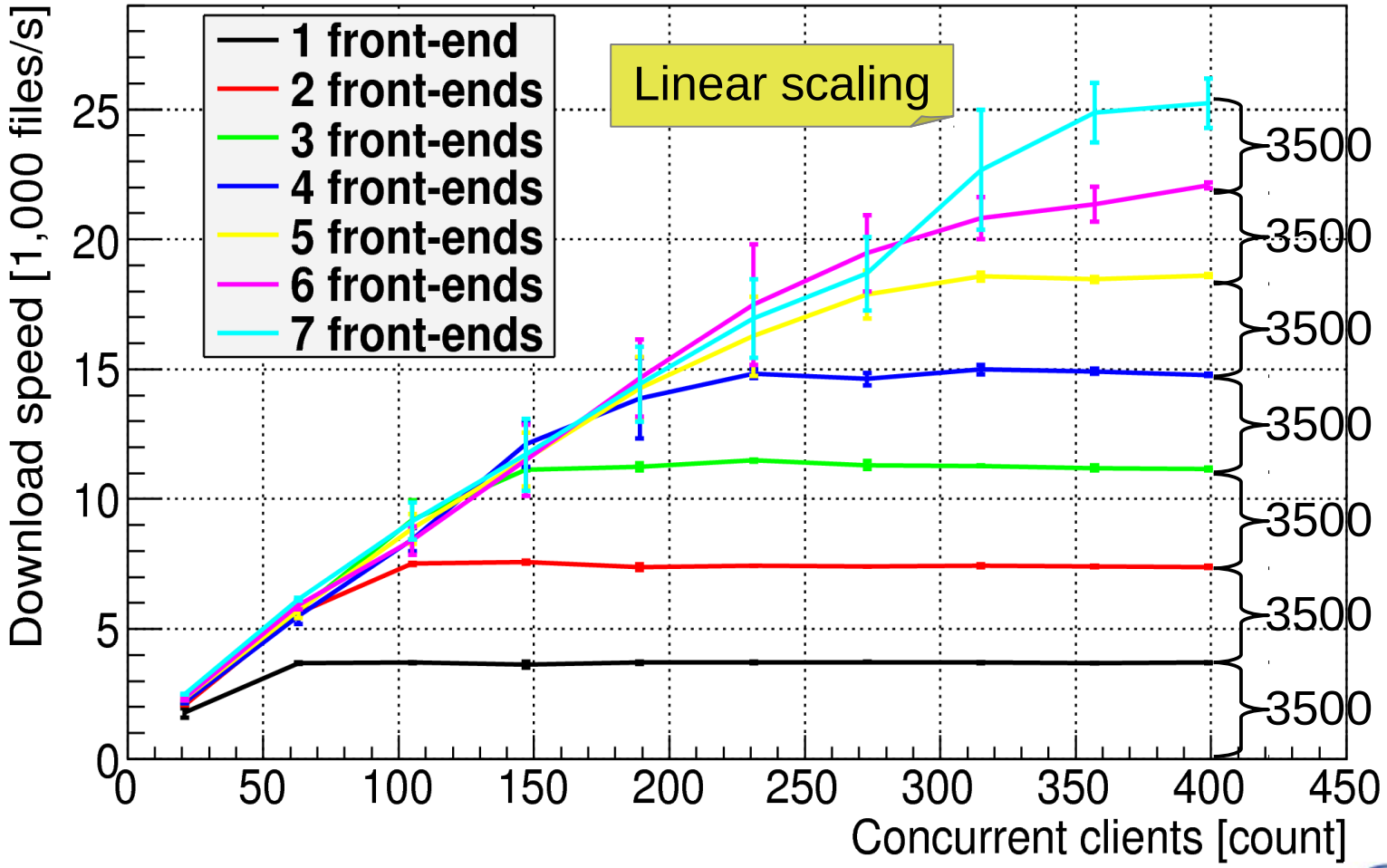
Front-end scalability



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

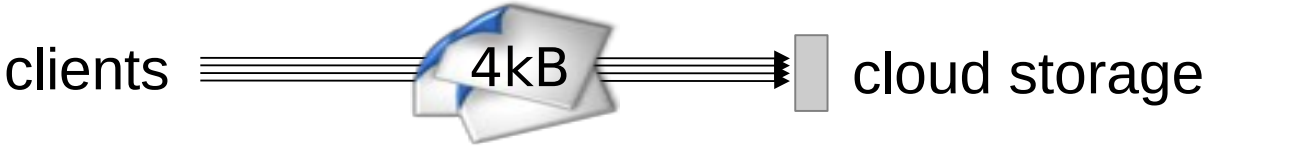
Front-end scalability

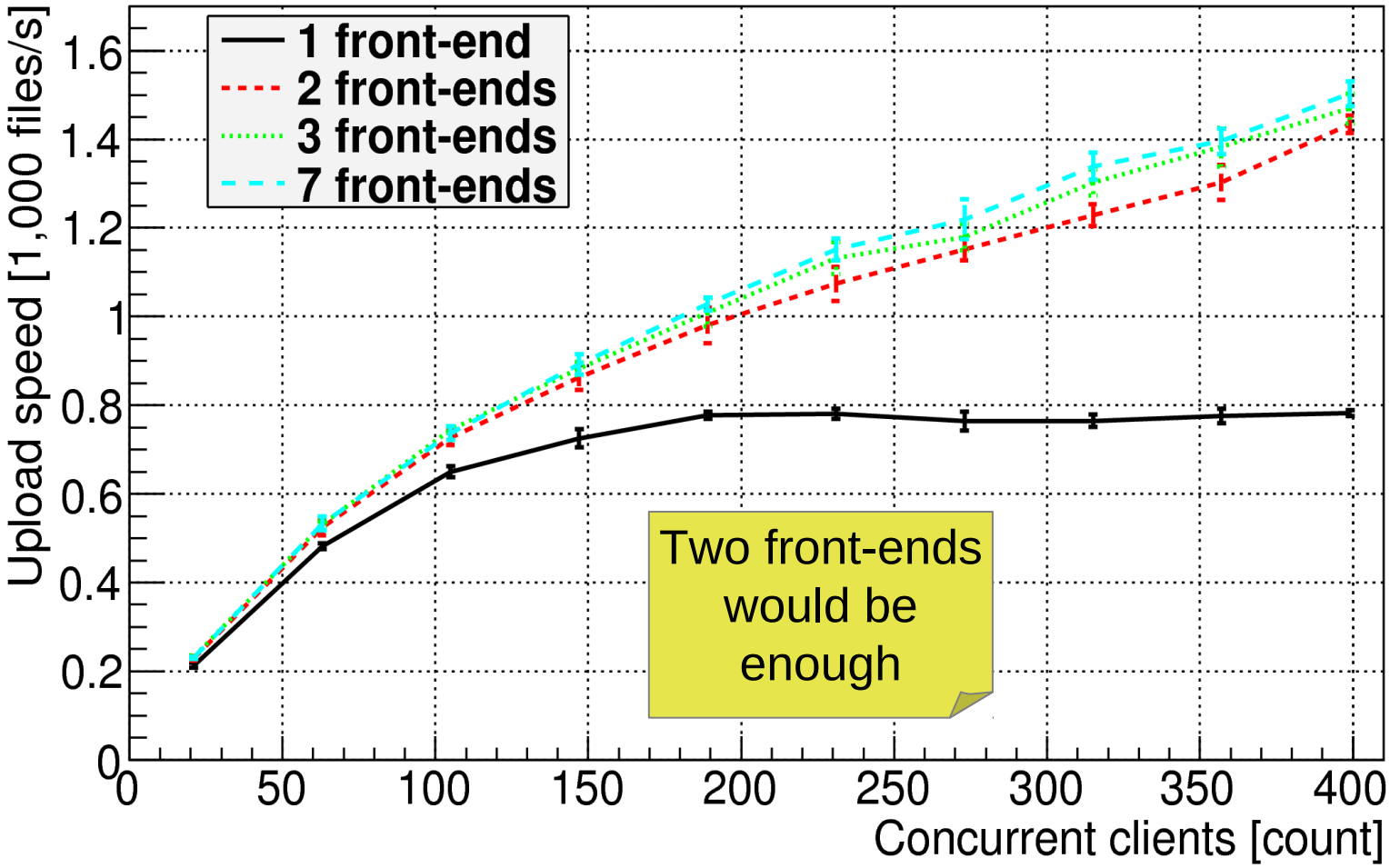




CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Front-end scalability





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

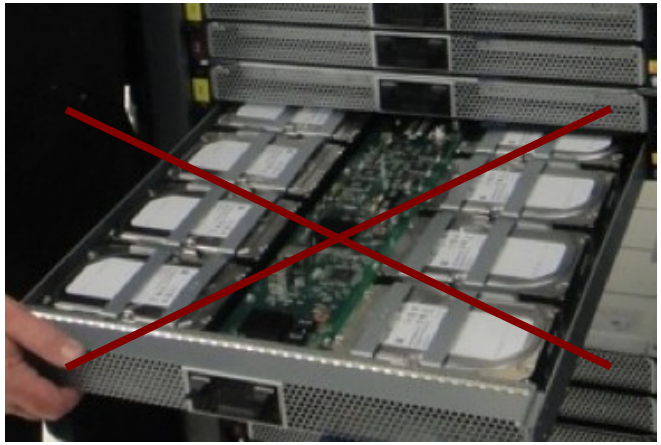
Raw performance summary

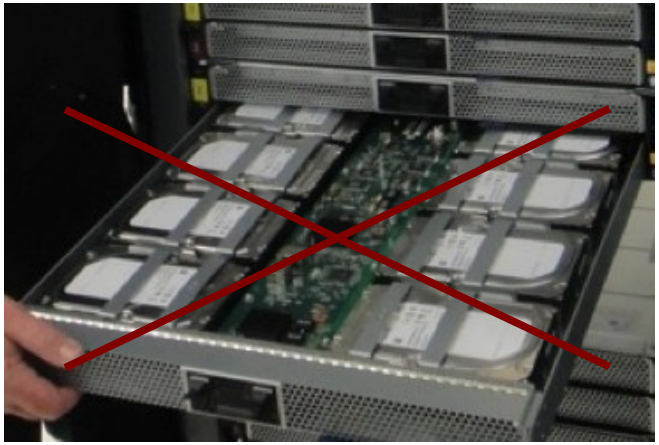
- Metadata (4kB) performance
  - 2,500 files/second upload
  - 25,000 files/second download
- Throughput (100MB) performance
  - 20Gbit network fully utilized
- Front-end scalability
  - Each front-end can download 3500 files/s
  - Each front-end can upload 550 MB/s

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

Recovery after powering off a chassis

Two blades are powered off:



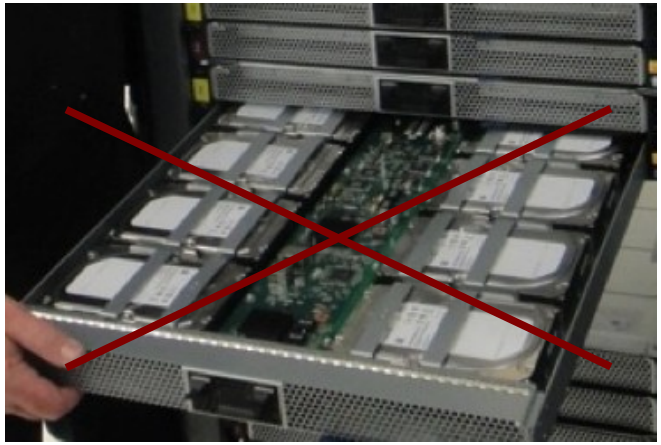


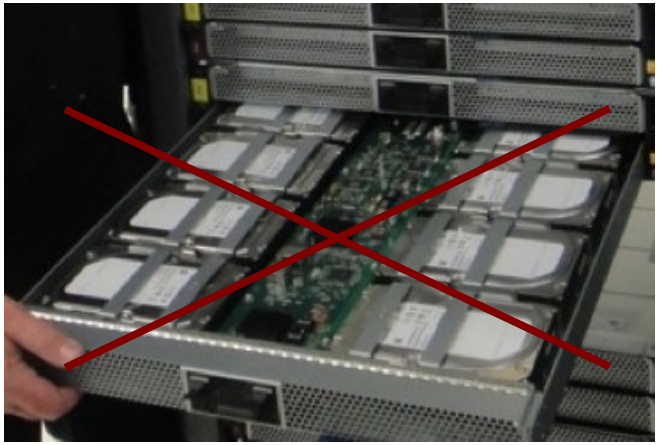
16 disks down

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

Recovery after powering off a chassis

Two blades are powered off:



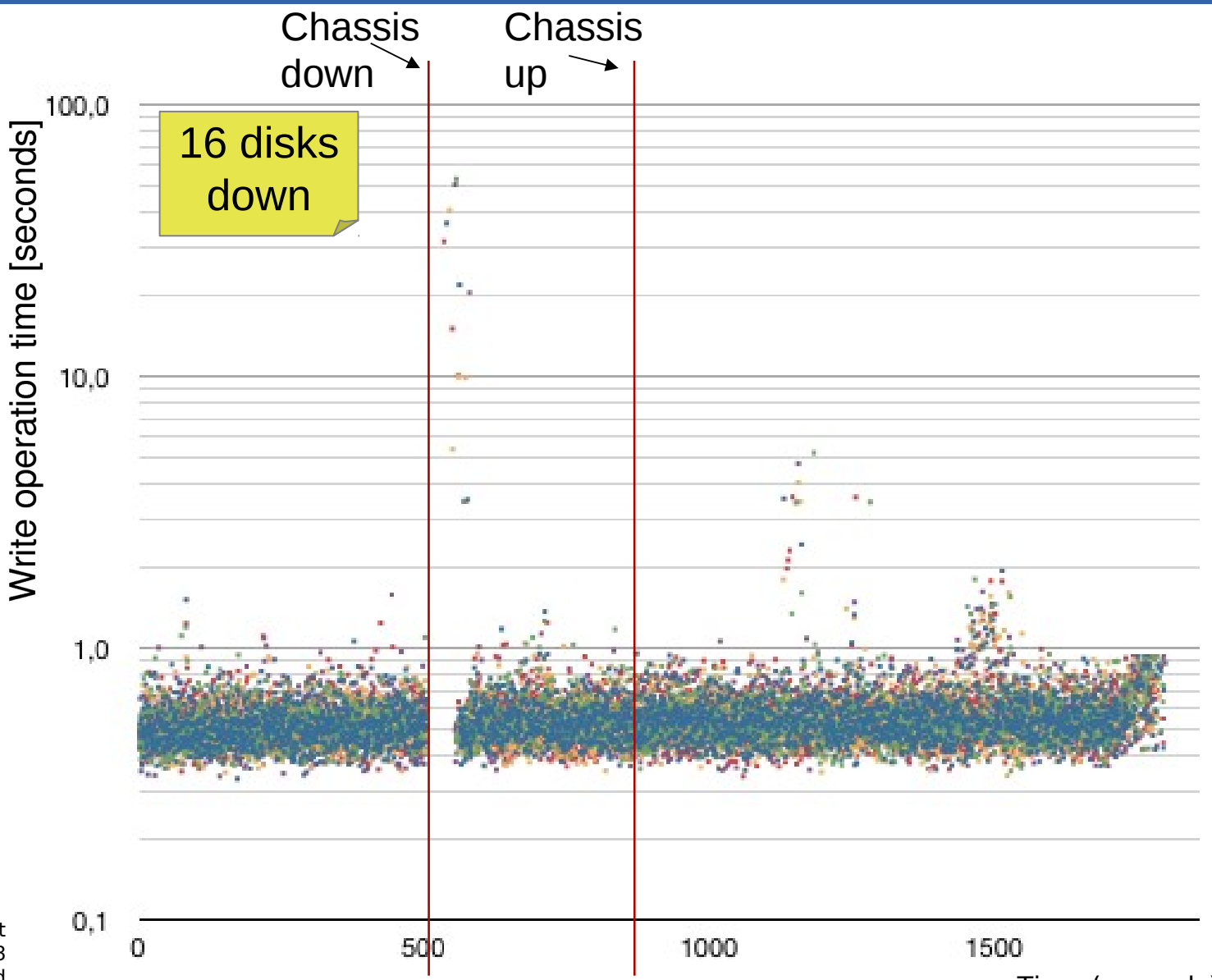


16 disks down

Uploads and downloads continue normally?

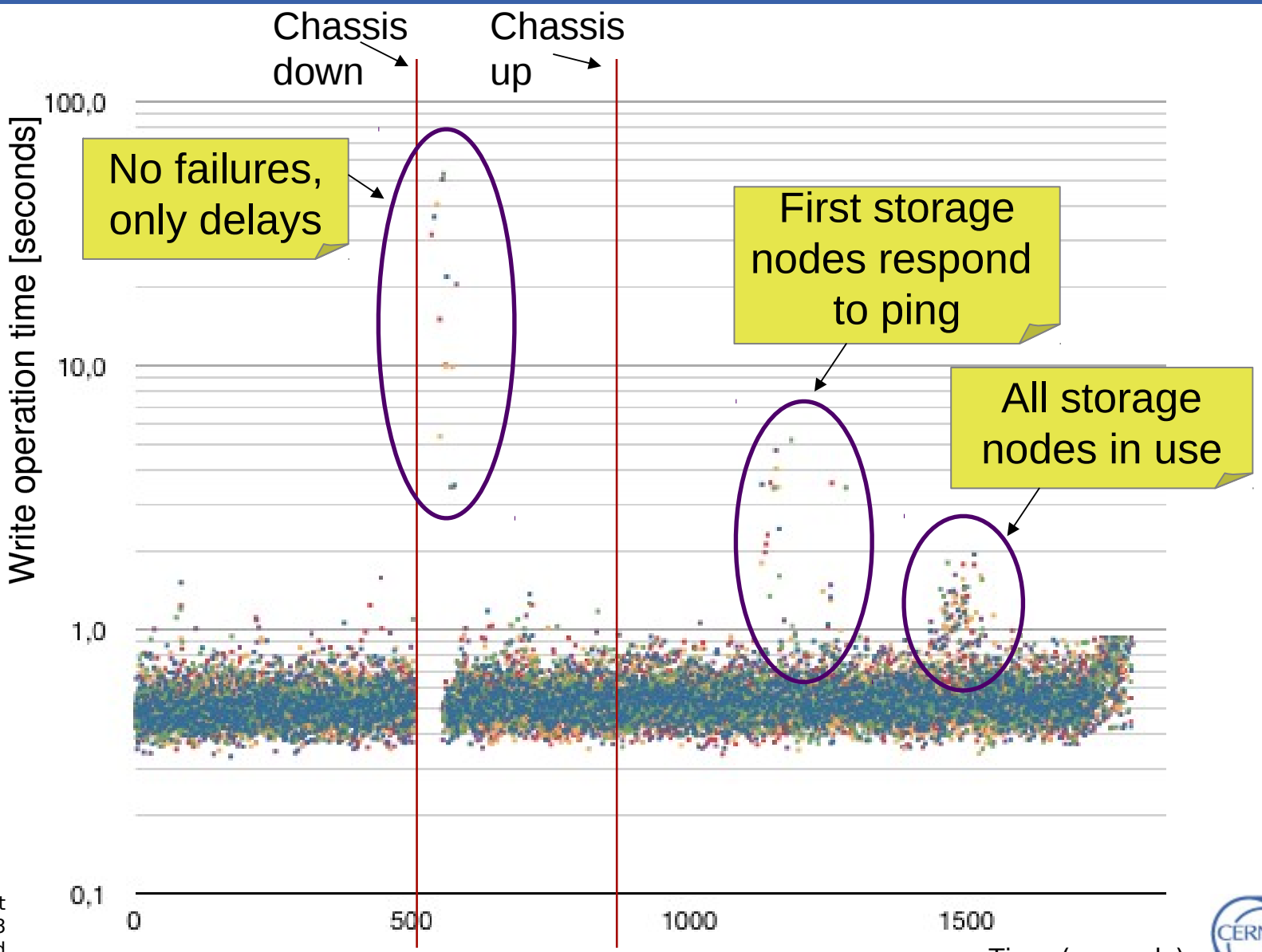
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Recovery after powering off a chassis



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

Recovery after powering off a chassis



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

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

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

File system with Huawei back-end

- Implementation

CernVM File system

  - 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 with Huawei back-end

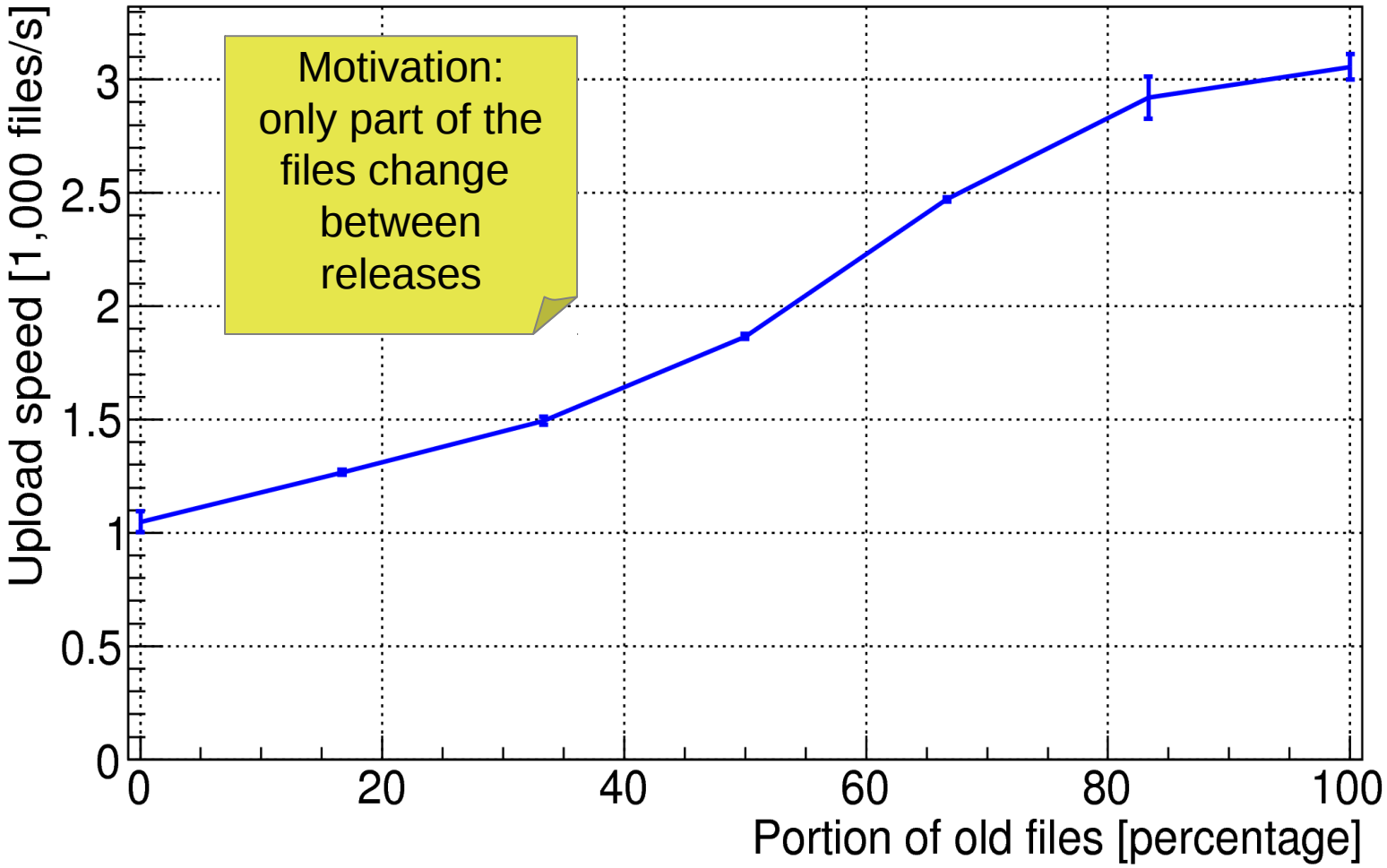
- Implementation
  - Files are uploaded to multiple buckets in the cloud storage
  - Files are downloaded with unified name space
    http://cloud.cern.ch/~~bucket-42~~/file001.bin
    http://cloud.cern.ch/file001.bin
- Result
  - Full publish procedure tested to work using 30,000 small files
  - Upload speed 1200 files/second (with 240 threads)

CernVM File system

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

CVMFS incremental uploads

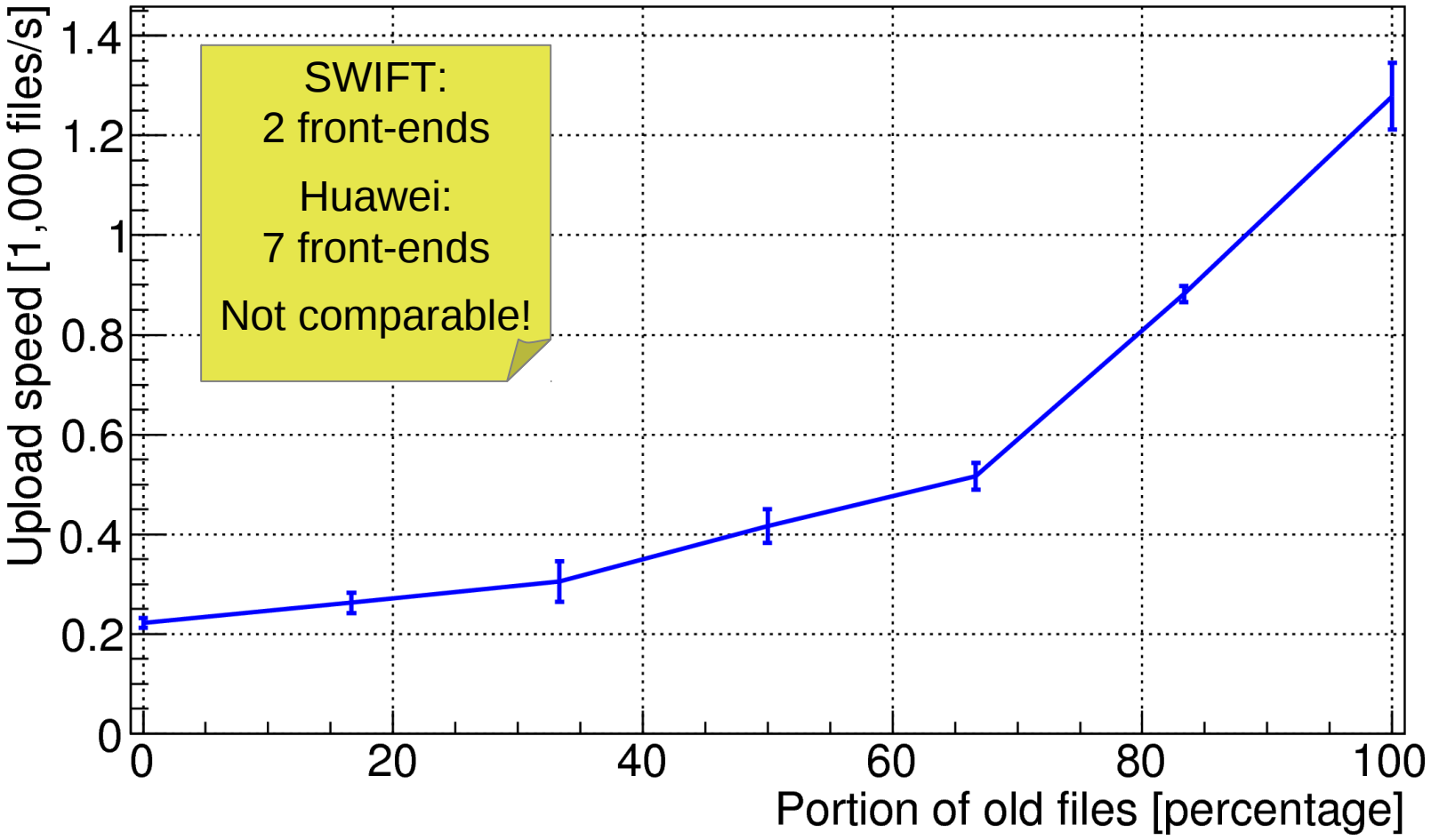
Uploading 30,000 files (of average size 10kB)
to Huawei UDS



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

CVMFS incremental uploads

Uploading 30,000 files (of average size 10kB)
to OpenStack SWIFT



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

Conclusion

- 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

Future plans

- Short term
  - Benchmark CVMFS with real release data
  - Test ROOT's new S3 plugin performance
- Long term
  - Second petabyte system with enterprise disks expected to arrive soon
  - Upgrade old Huawei cloud storage software version
  - Replication tests between two 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's new S3 plugin performance
- Long term
  - Second petabyte system with enterprise disks expected to arrive soon
  - Upgrade old Huawei cloud storage software version
  - Replication tests between two cloud storages
  - Prove total cost of ownership (TCO) gains of the system as part of a production service

Thank you!

seppo.heikkila@cern.ch

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