DSS

Data & Storage Services

CERN IT Department

Huawei Cloud Storage

Maitane Zotes Resines, CERN IT
(With contributions from Wang Lu, IHEP Beijing)

Openlab Major Review Meeting
31. January 2013
CERN, Geneva

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

HUAWEI

CERN openlab

Outline

- Timeline
- Huawei setup and benchmark
- Last testing phase's results
- OSC scalability test
- 21 Client uploads and downloads
- Recovery after powering off a chassis test
- Long term stability test
- No cache downloads
- IHEP status
- Conclusions and future plans

Timeline

1 Year of Huawei...

Lightning Talk

IHEP Setup

Major Review

20 Client testing

Major Review

Project starts

5 Client testing

Minor Review

January 2012

January 2013

First tests

Upgrade of the system

Major Review

Minor Review

21 Client testing

Mounting of the system

Board of Sponsors

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

DSS Huawei Setup

CERN IT Department

CERN IT Departmen
CH-1211 Genève 2
Switzerlan
www.cern.ch/i

Huawei Setup

DSS

CERN IT Department

CERN IT Departmen
CH-1211 Genève 2
Switzerlan
www.cern.ch/i

DSS

Huawei Setup

CERN IT Department

CERN IT Departmen
CH-1211 Genève 2
Switzerlan
www.cern.ch/i

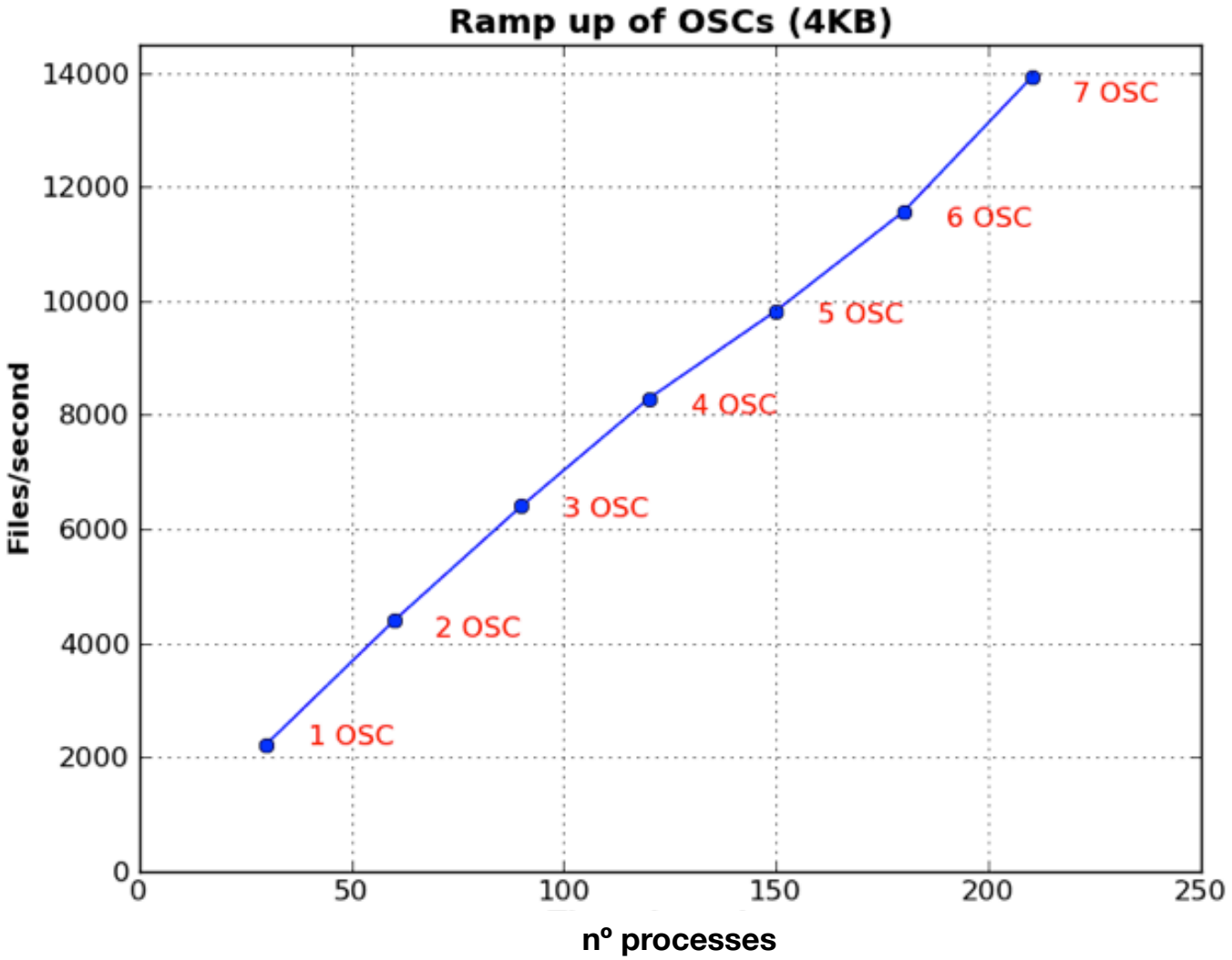
Benchmark

- C++ framework integrated with ROOT
  - Integrated with the Python benchmark
  - ssh connection to clients using lxplus
  - Histograms about specific metrics
    - Operation time
    - Tx and receive rate
    - CPU and memory utilization
- S3 module
  - Amazon S3 aws library for Python

Last phase results

- 5 Client uploads
  - 383 files/second for the metadata scaling
  - Fully filled 5Gb bandwidth limit
- 5 Client downloads
  - 8500 files/second for the metadata scaling
  - Fully filled 5Gb bandwidth limit
- 20 Client downloads
  - 18000 files/second for the metadata scaling
  - Reached 18Gb network bandwidth

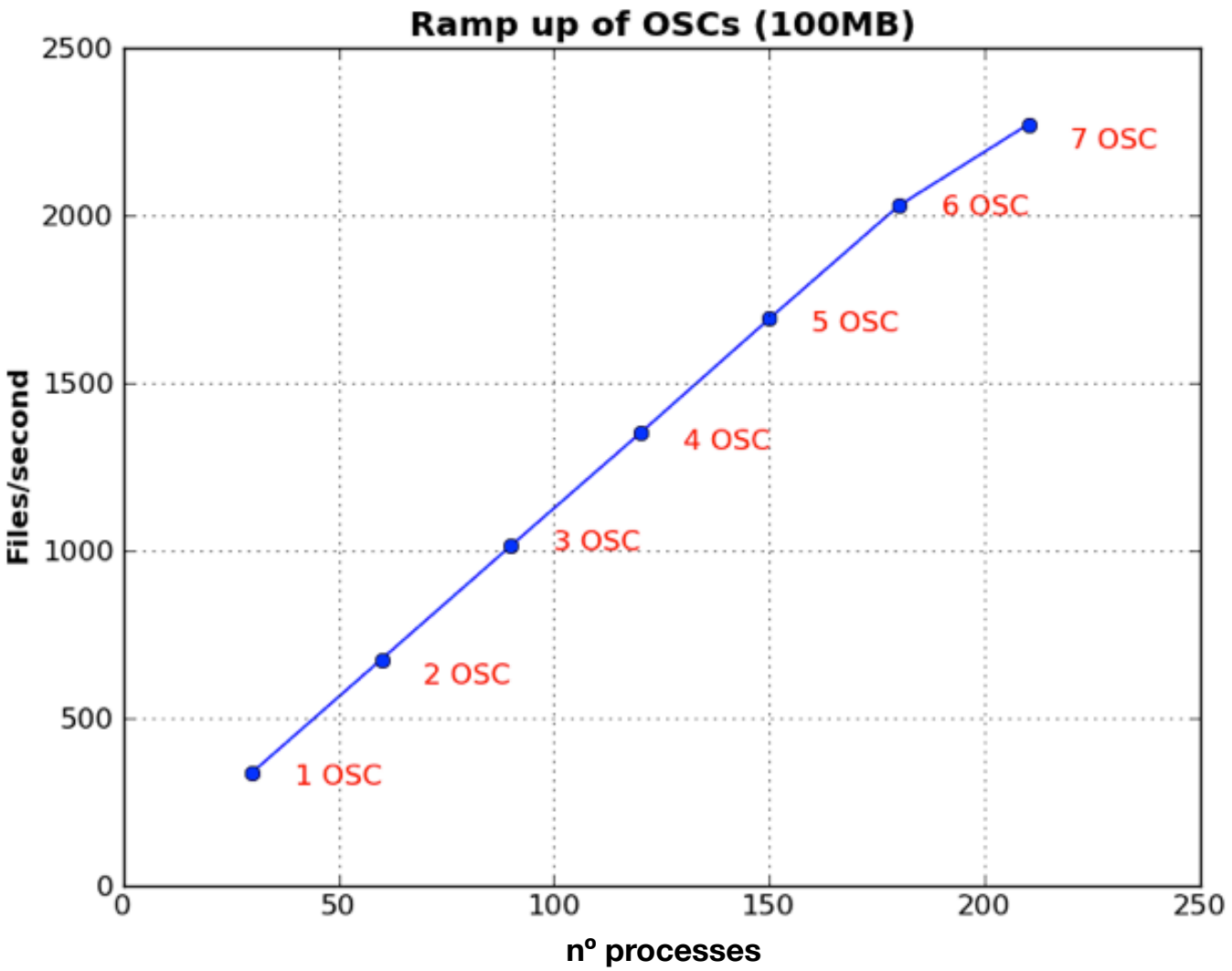
OSC scalability



Linear scaling as each OSC processes around 2000 files/sec

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

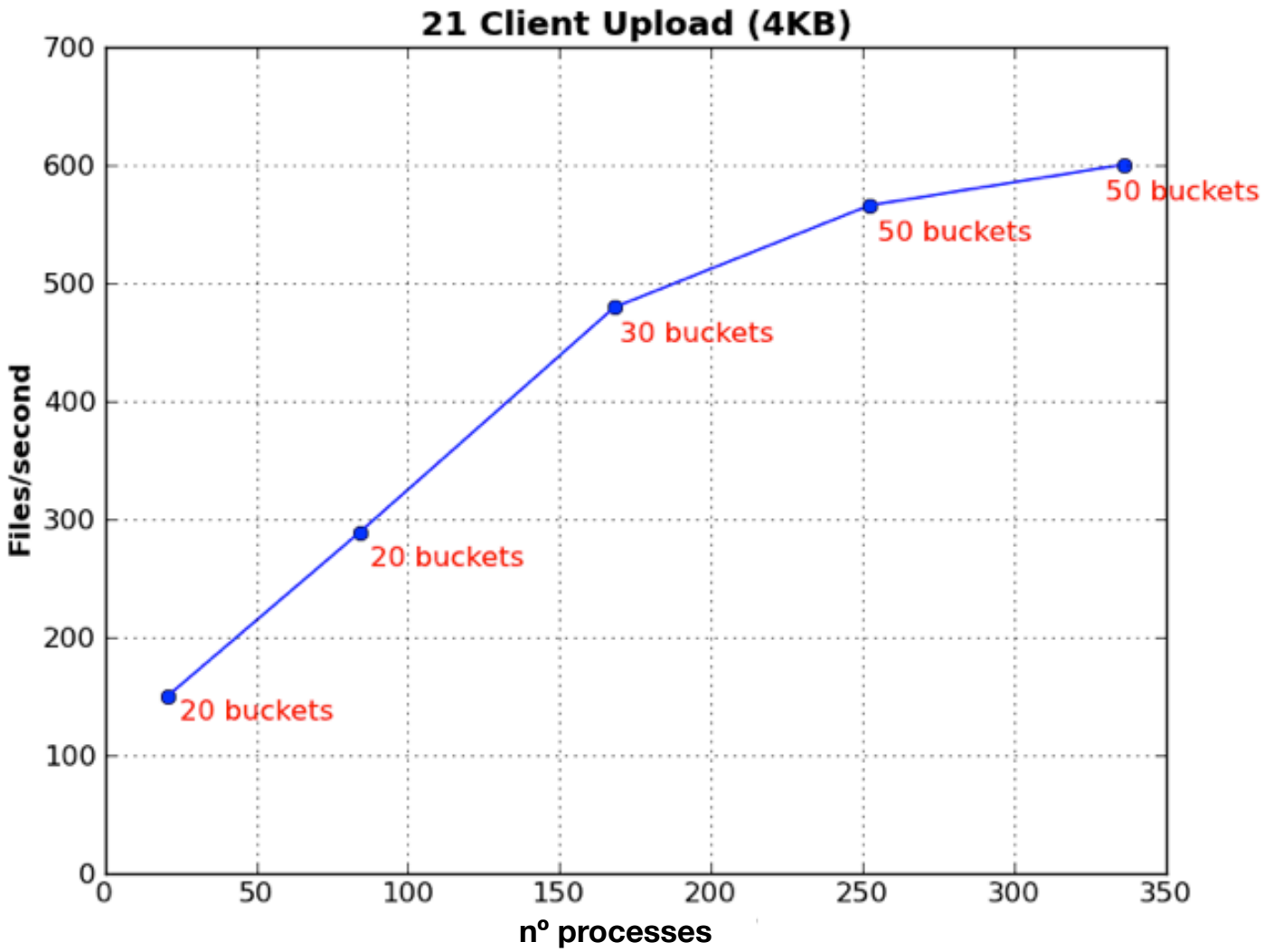
OSC scalability



Each client has 1Gb network bandwidth, so with 3 clients (10 threads/client), each OSC handles around 350MB/second

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

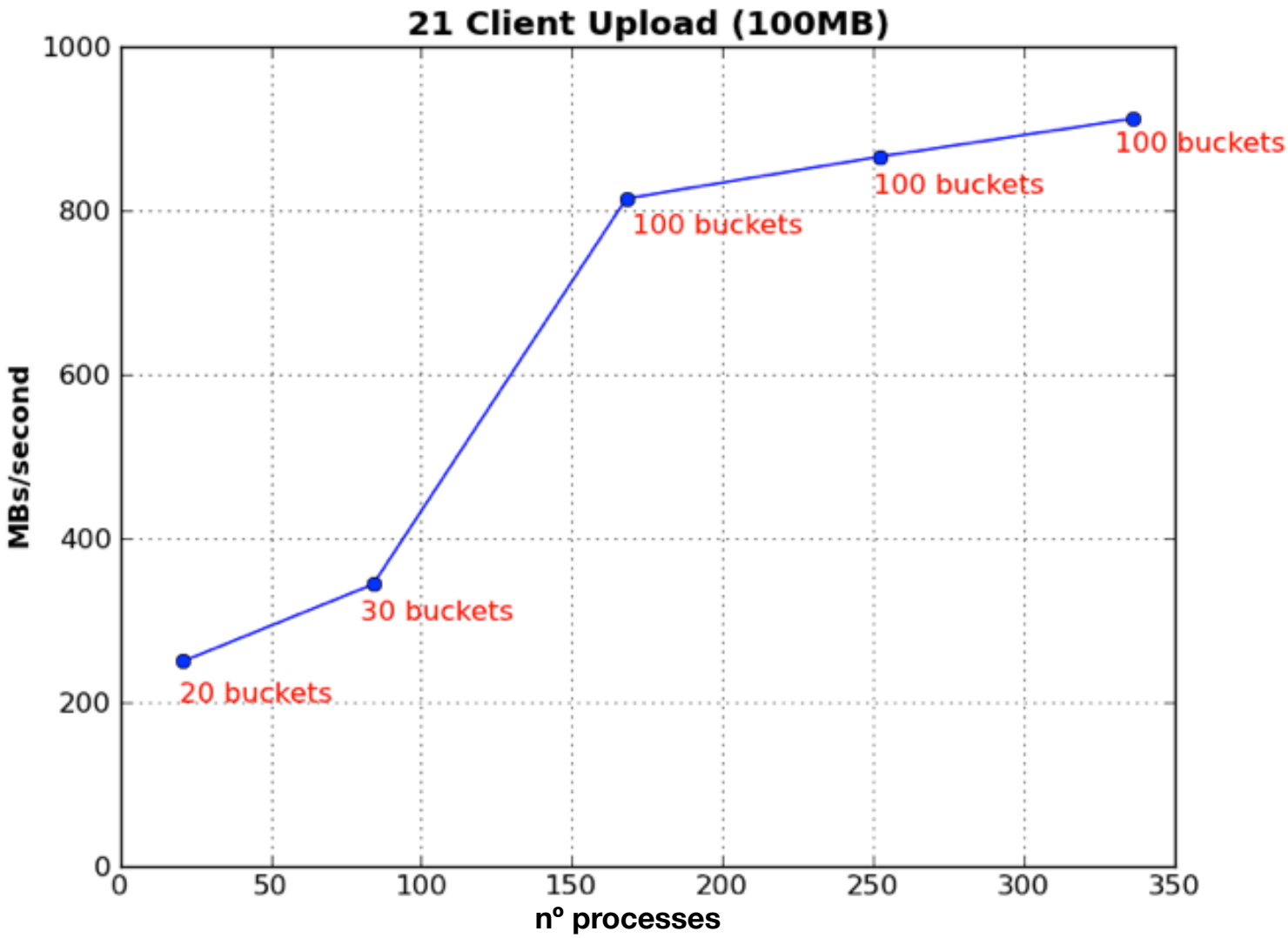
Upload Scaling



Performance variable depending on the number of buckets used

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

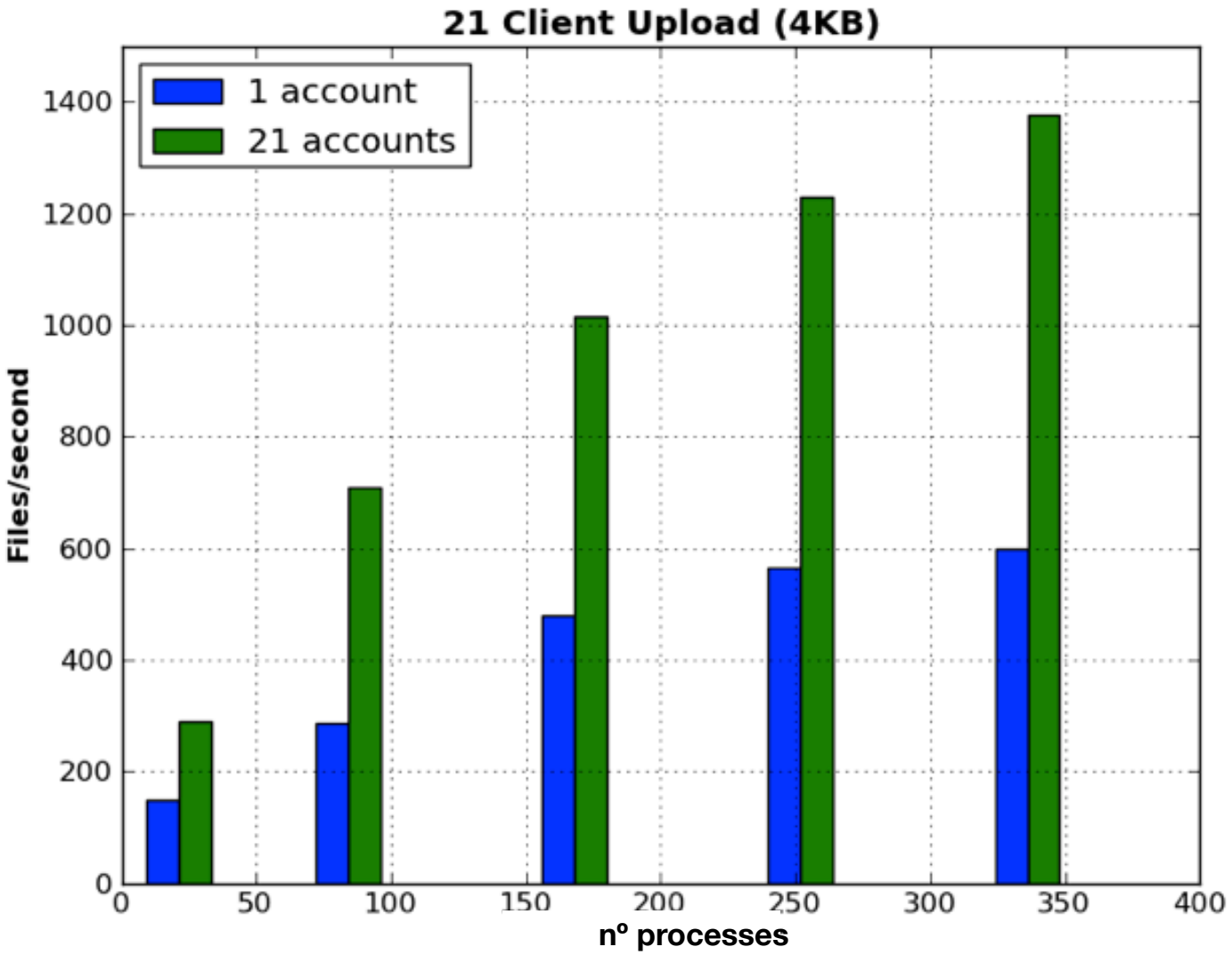
Upload Scaling



Performance variable depending on the number of buckets used

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

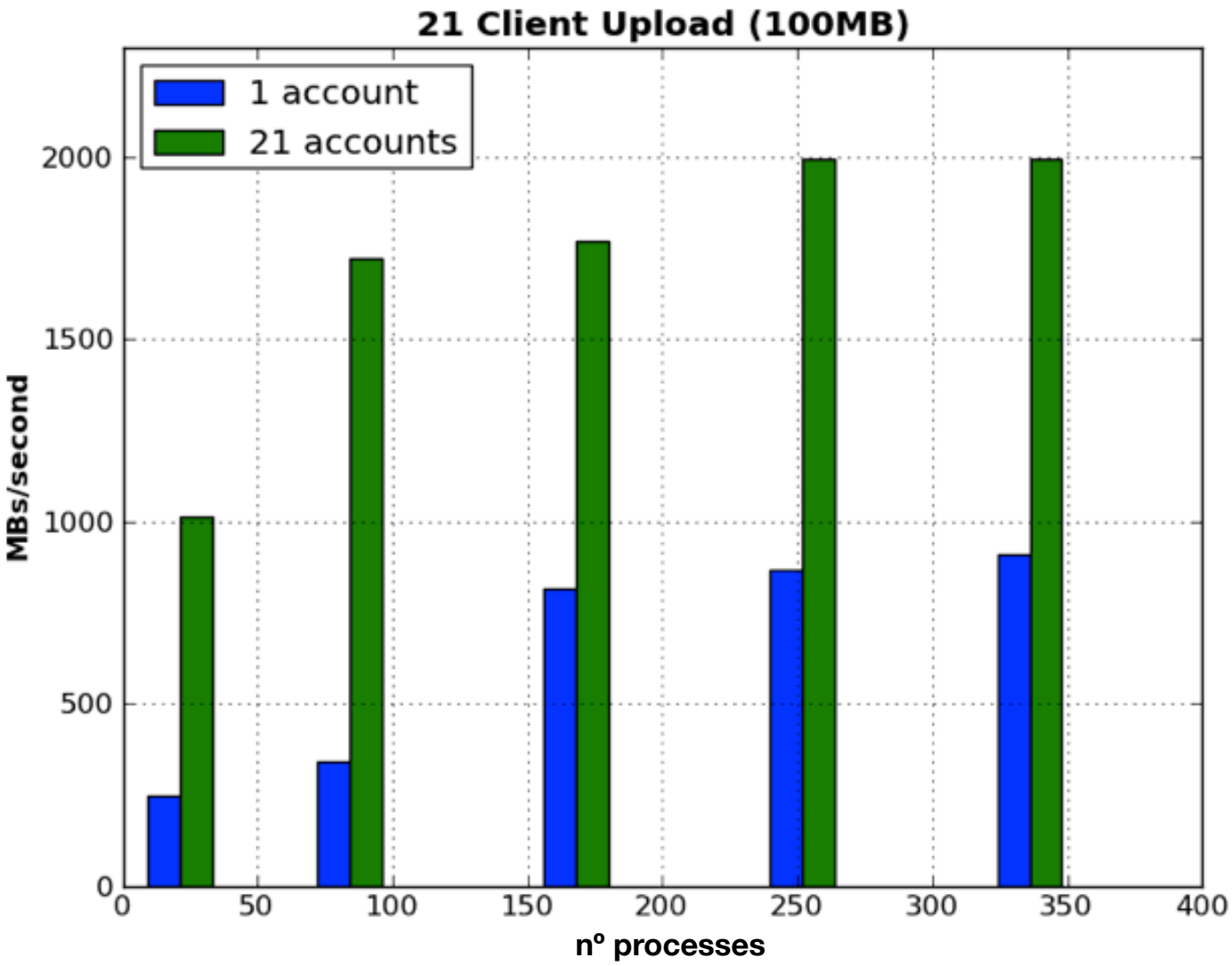
Upload Scaling: Multiple Accounts



21 account test doubles the performance. Up to 1400 files/sec

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

Upload Scaling: Multiple Accounts



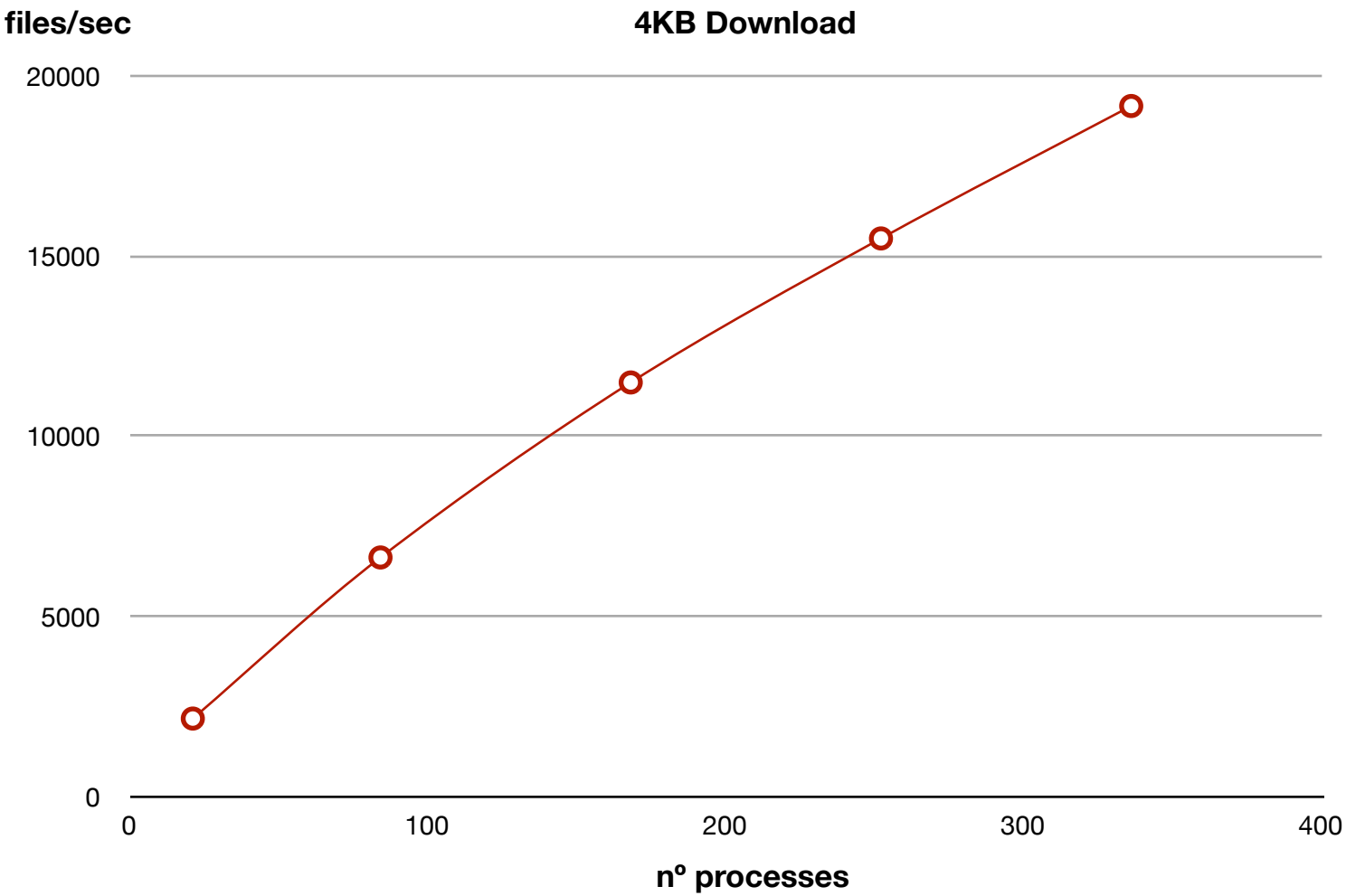
18Gb bandwidth limit reached using 21 accounts

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

Upload Scaling: Conclusion

- Bucket number used is a bottleneck
- The more buckets used, the faster the operation
- Limit of 100 buckets/account
- 1 Account/client looks to reach bandwitdth limit
  - 21 accounts * 100 buckets = 2100 buckets

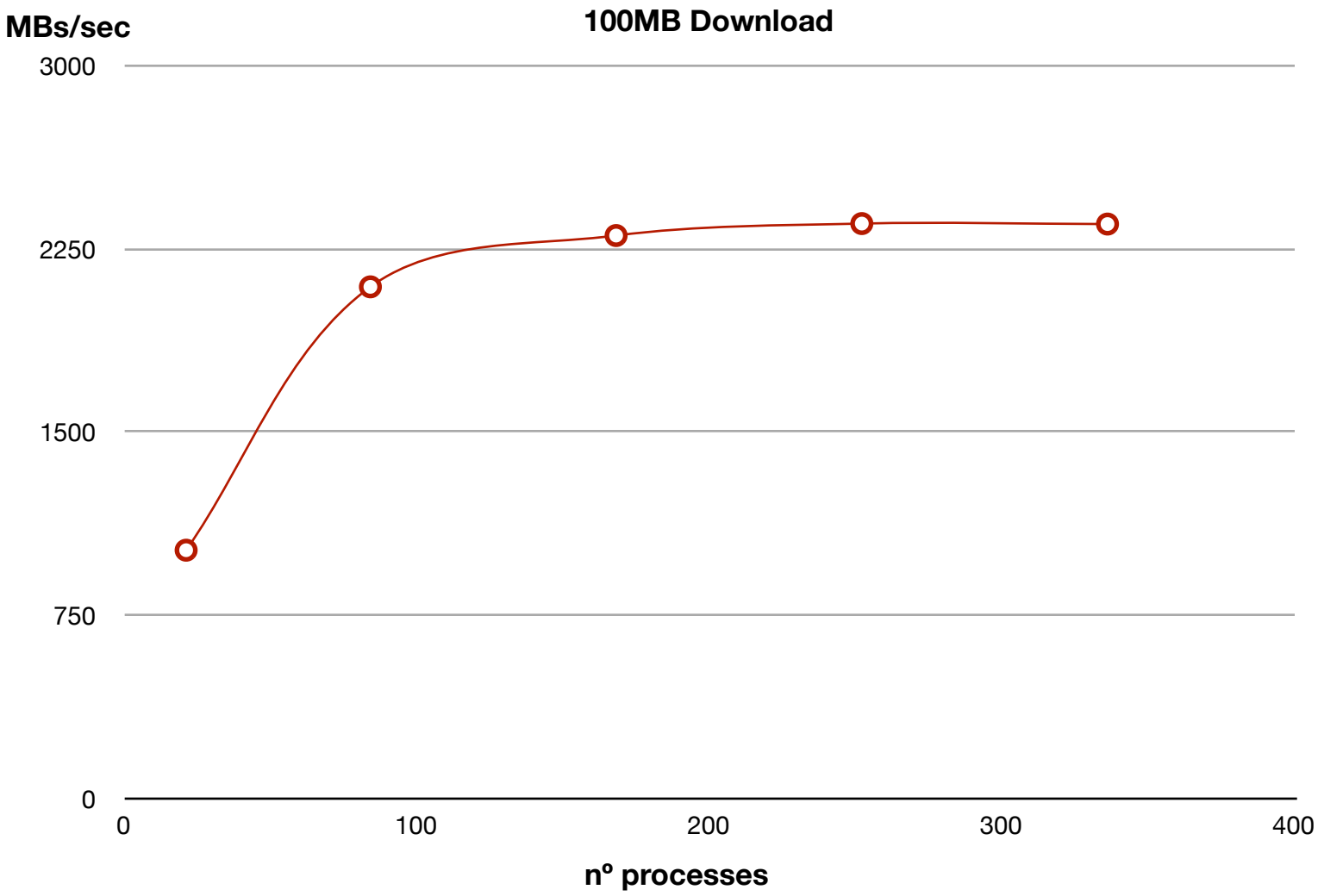
Download Scaling



Scales up to 18000 files/second

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

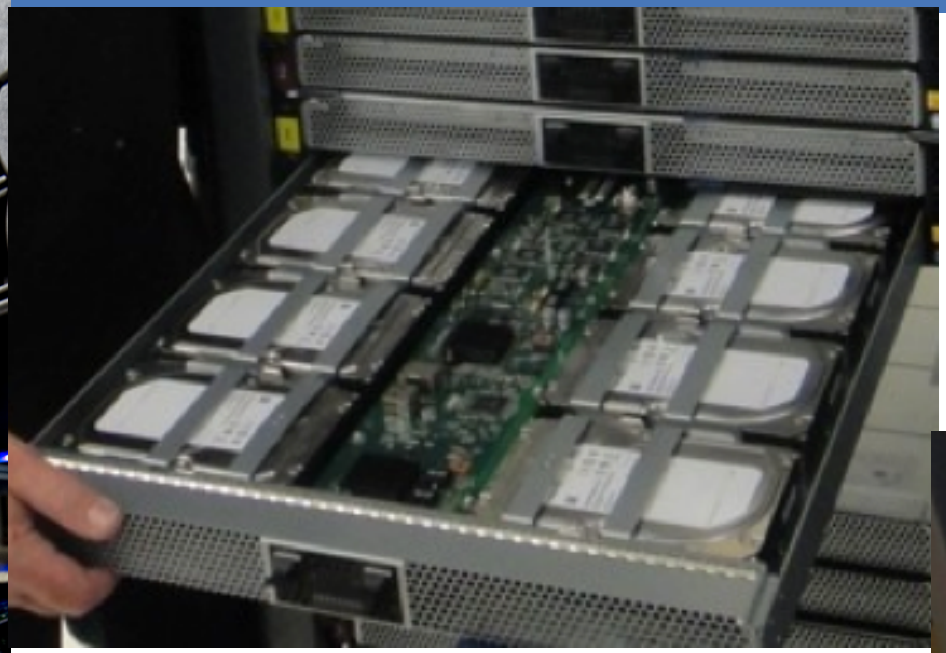
Download Scaling



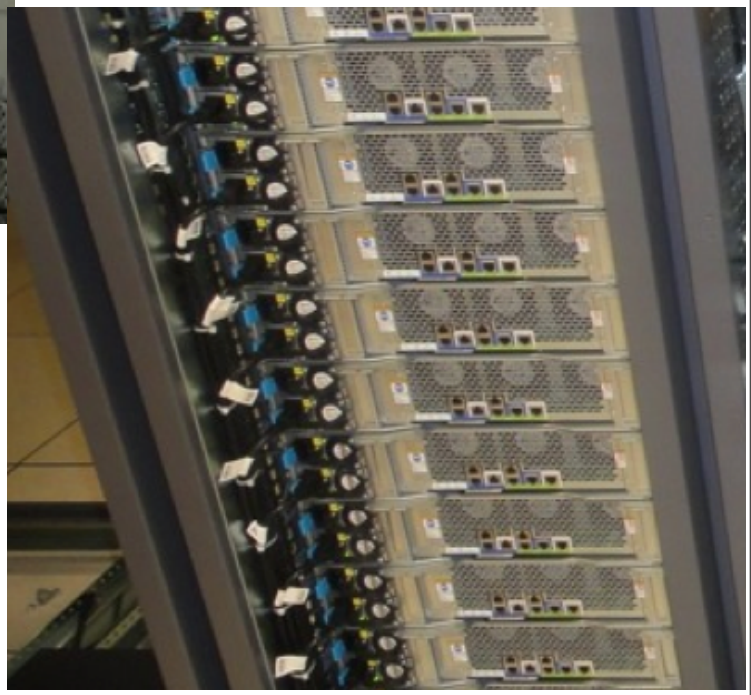
21 clients reach bandwidth limit of 18Gb

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

Recovery after powering off a chassis

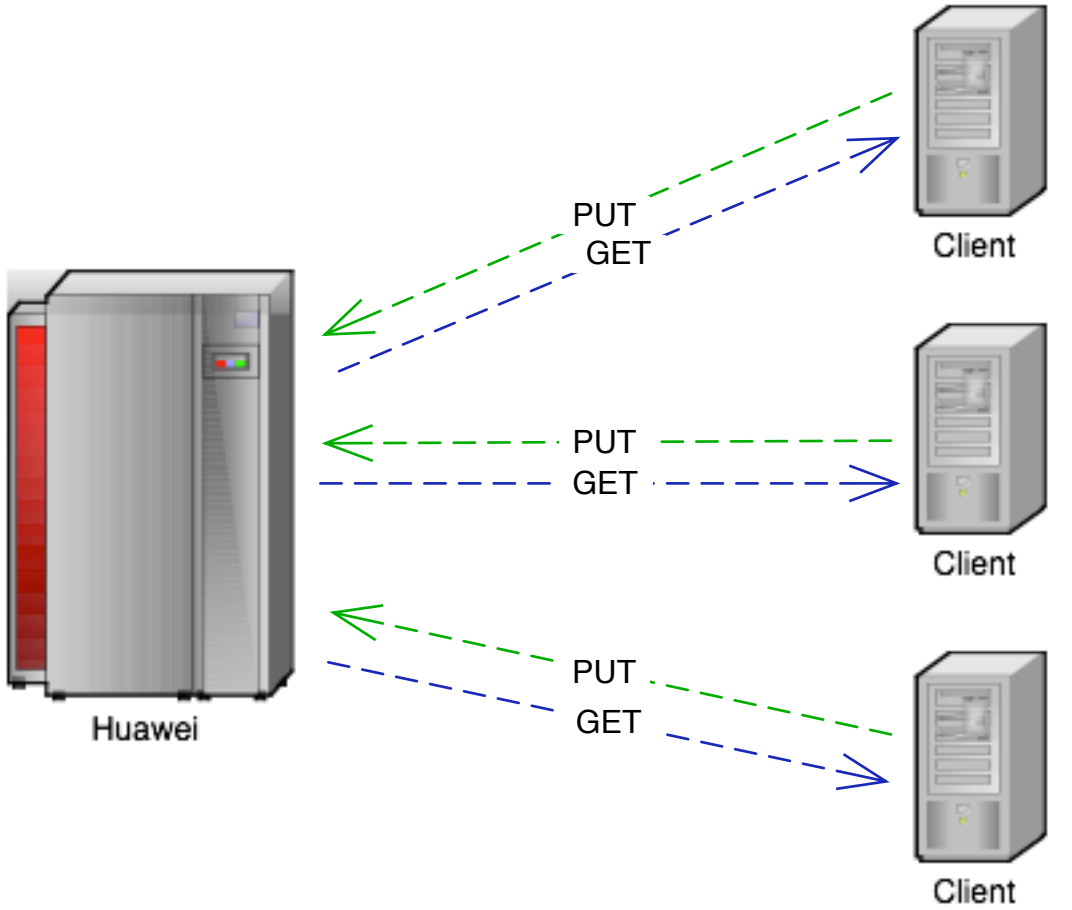


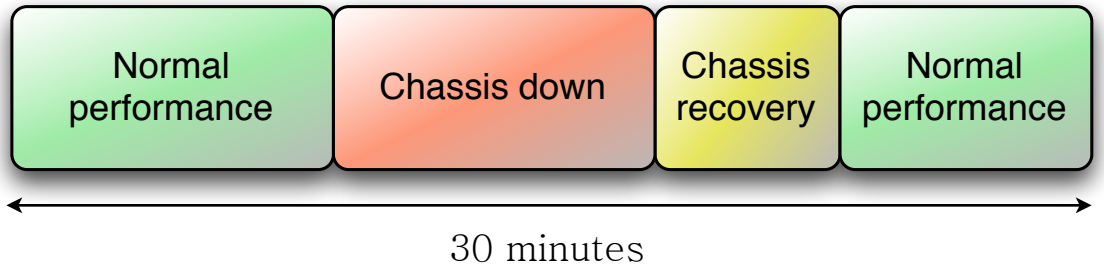
- Power it off by taking out the power cable



- 2 blades form a chassis
- Each blade, 8 disks
  - 16 disks will be powered off

Recovery after powering off a chassis





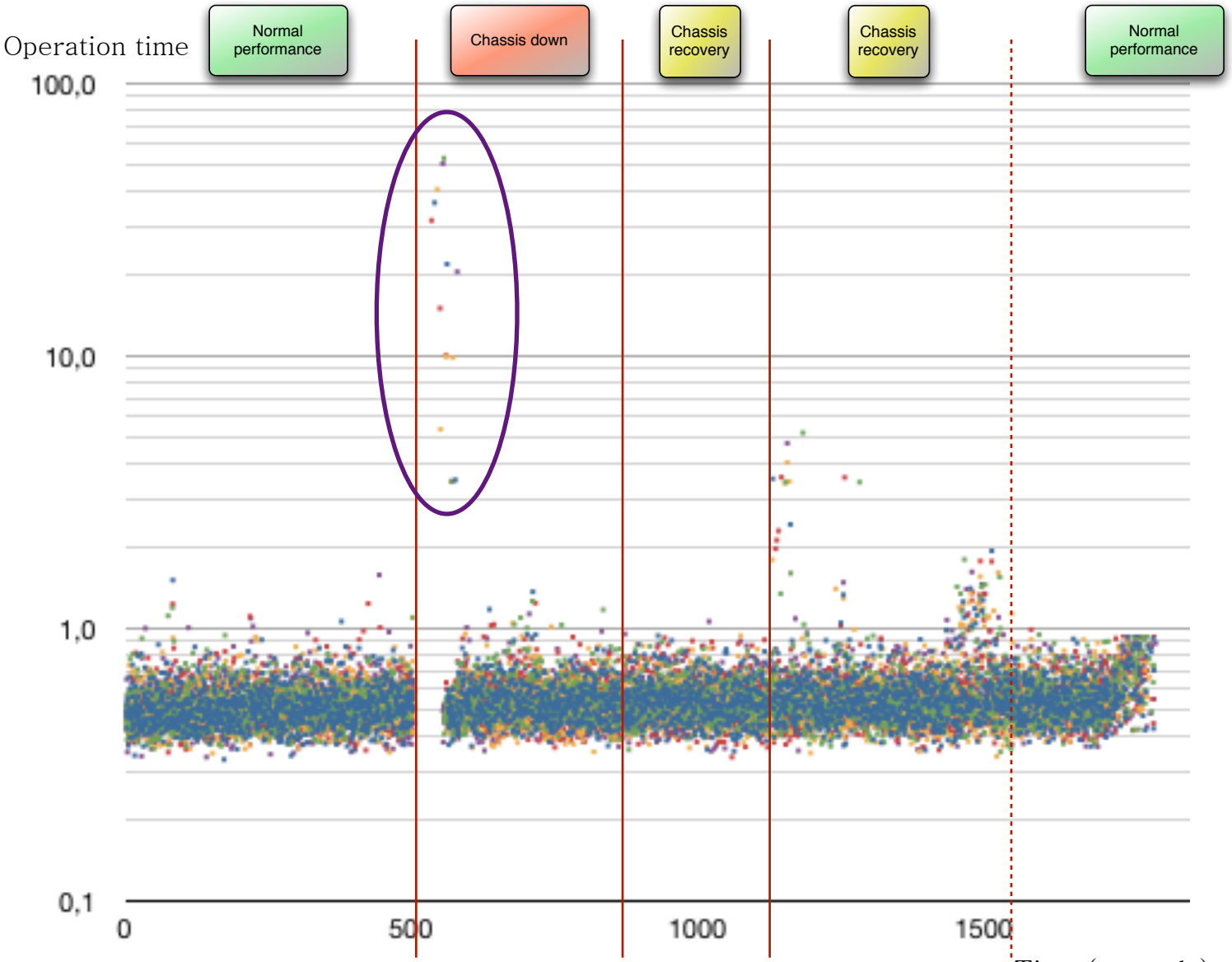
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Recovery after powering off a chassis: Write operation

CERN IT Department

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

Recovery after powering off a chassis: Write operation

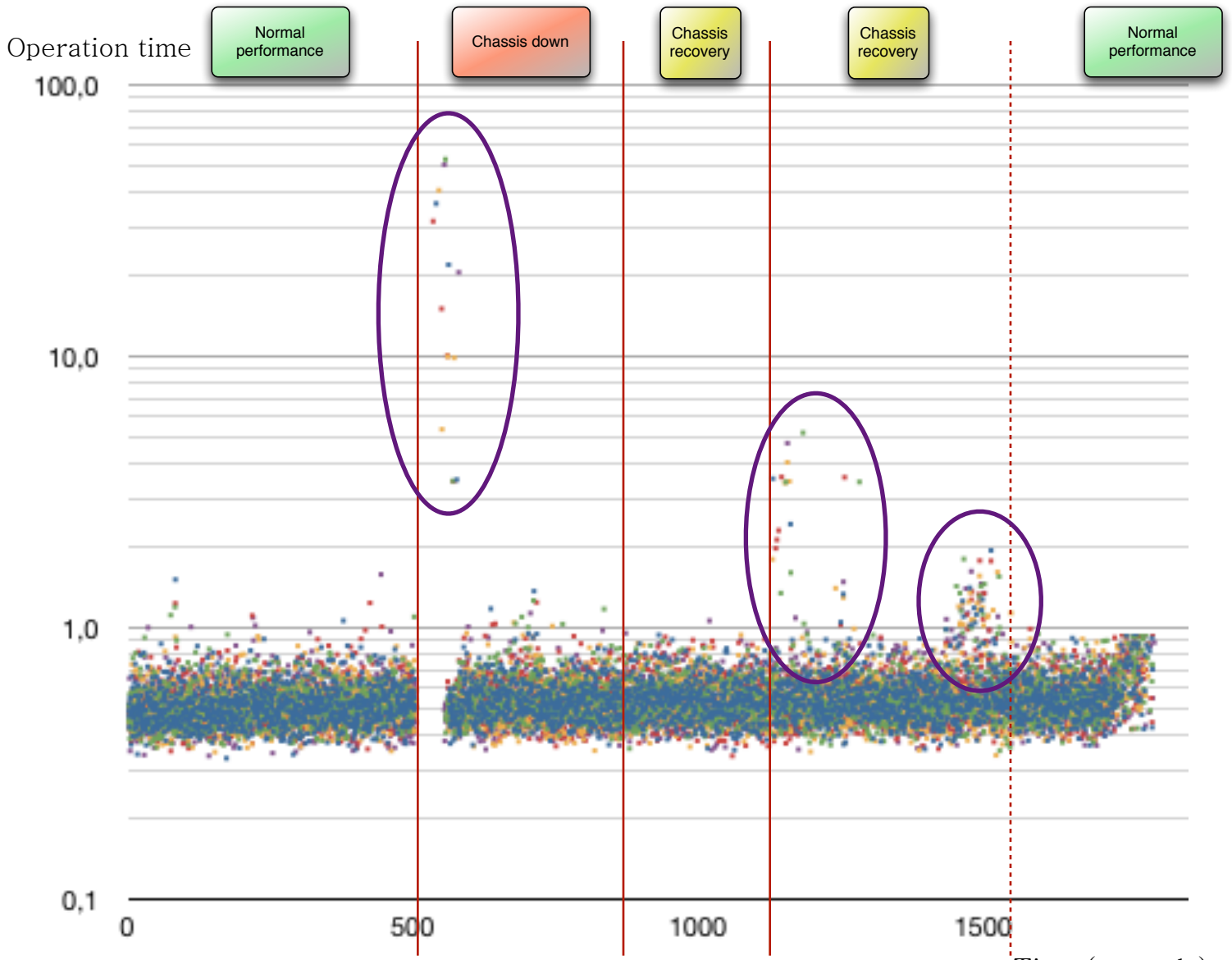


CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Recovery after powering off a chassis: Write operation

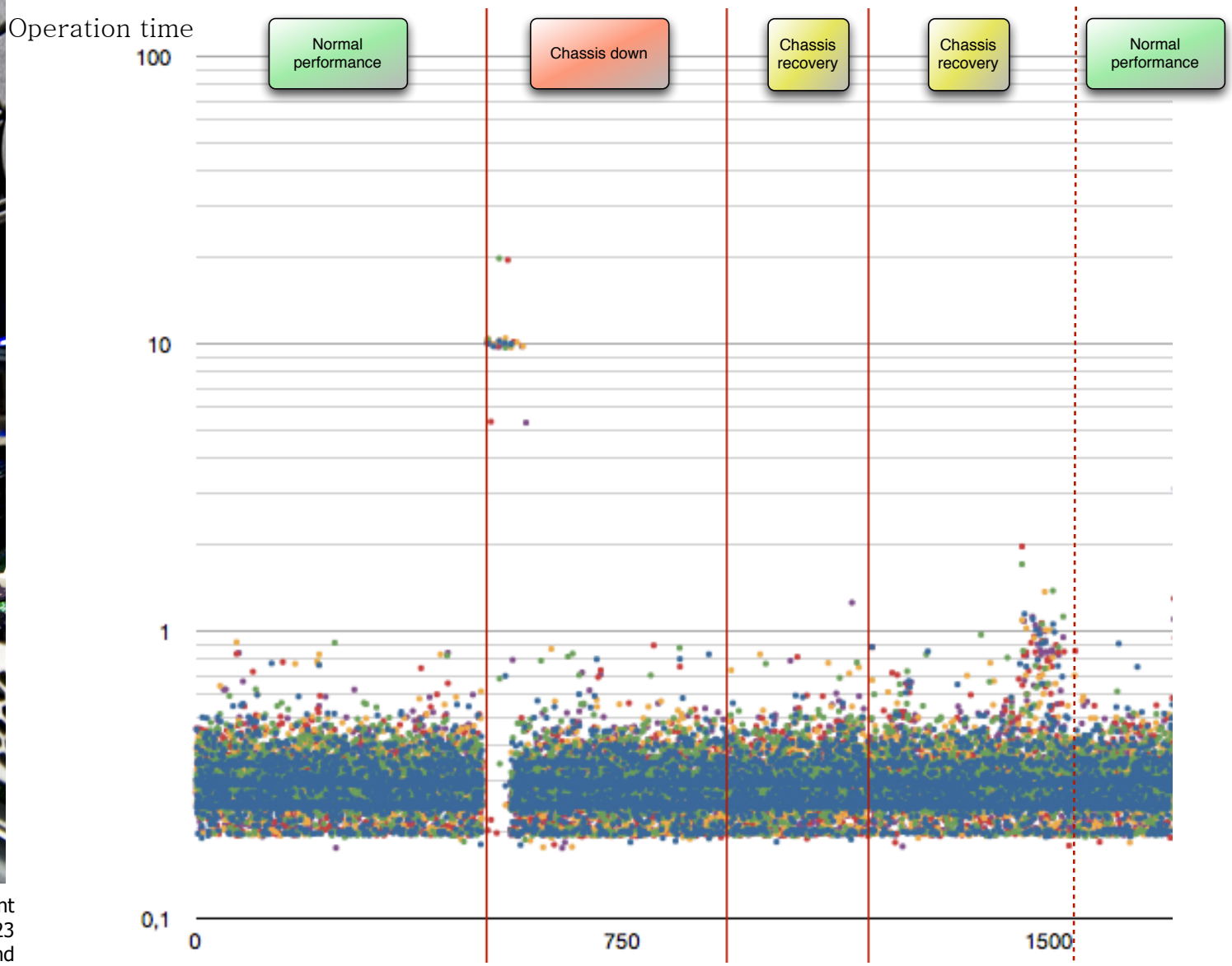
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Recovery after powering off a chassis: Write operation



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

Recovery after powering off a chassis: Read operation



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

Recovery after powering off a chassis: Read operation

Operation time

Normal performance

Chassis down

Chassis recovery

Chassis recovery

Normal performance

100

10

1

0,1

0

750

1500

Time (seconds)

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

Recovery after powering off a chassis: Read operation

Operation time

Normal performance

Chassis down

Chassis recovery

Chassis recovery

Normal performance

100

10

1

0,1

0

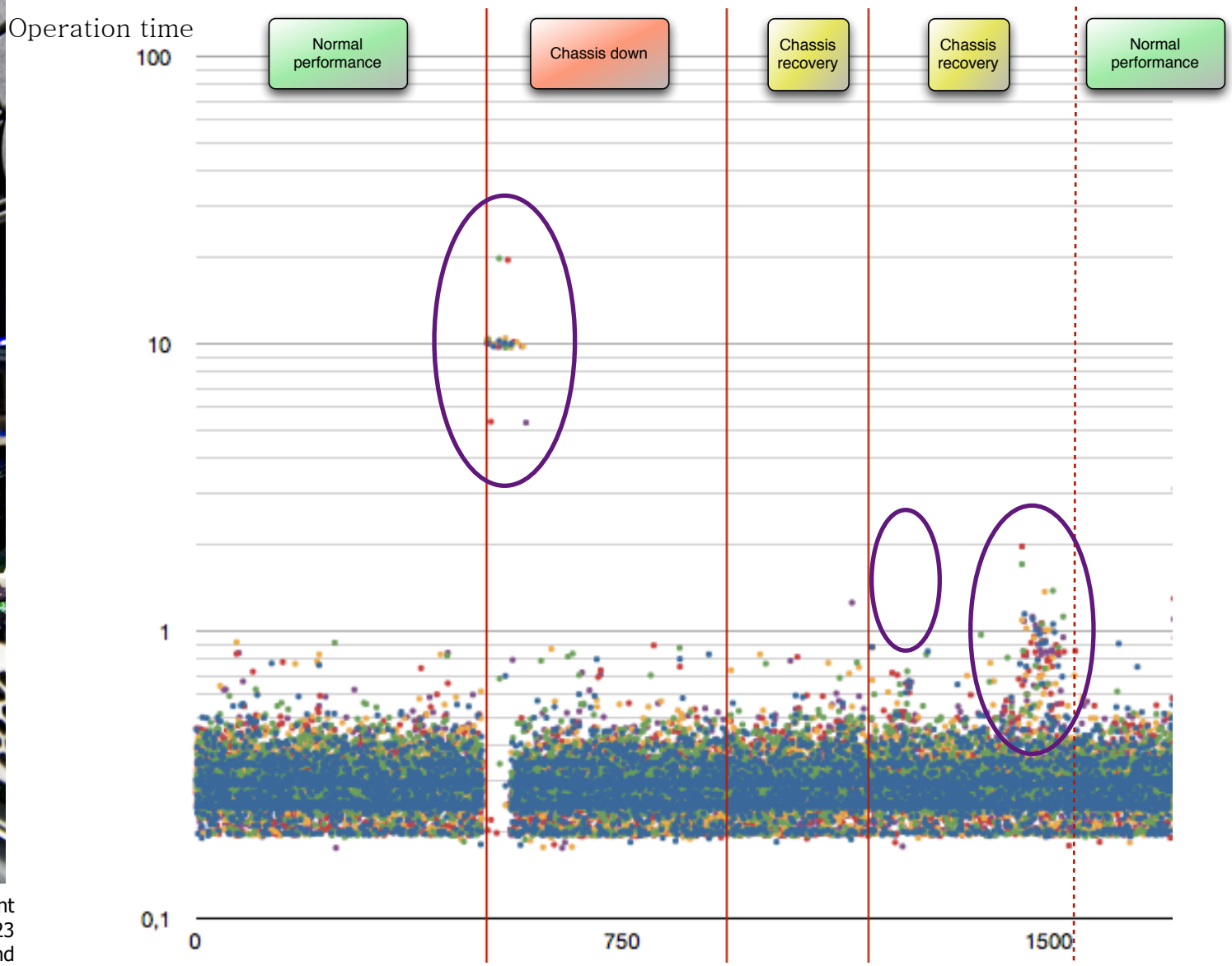
750

1500

Time (seconds)

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

Recovery after powering off a chassis: Read operation



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

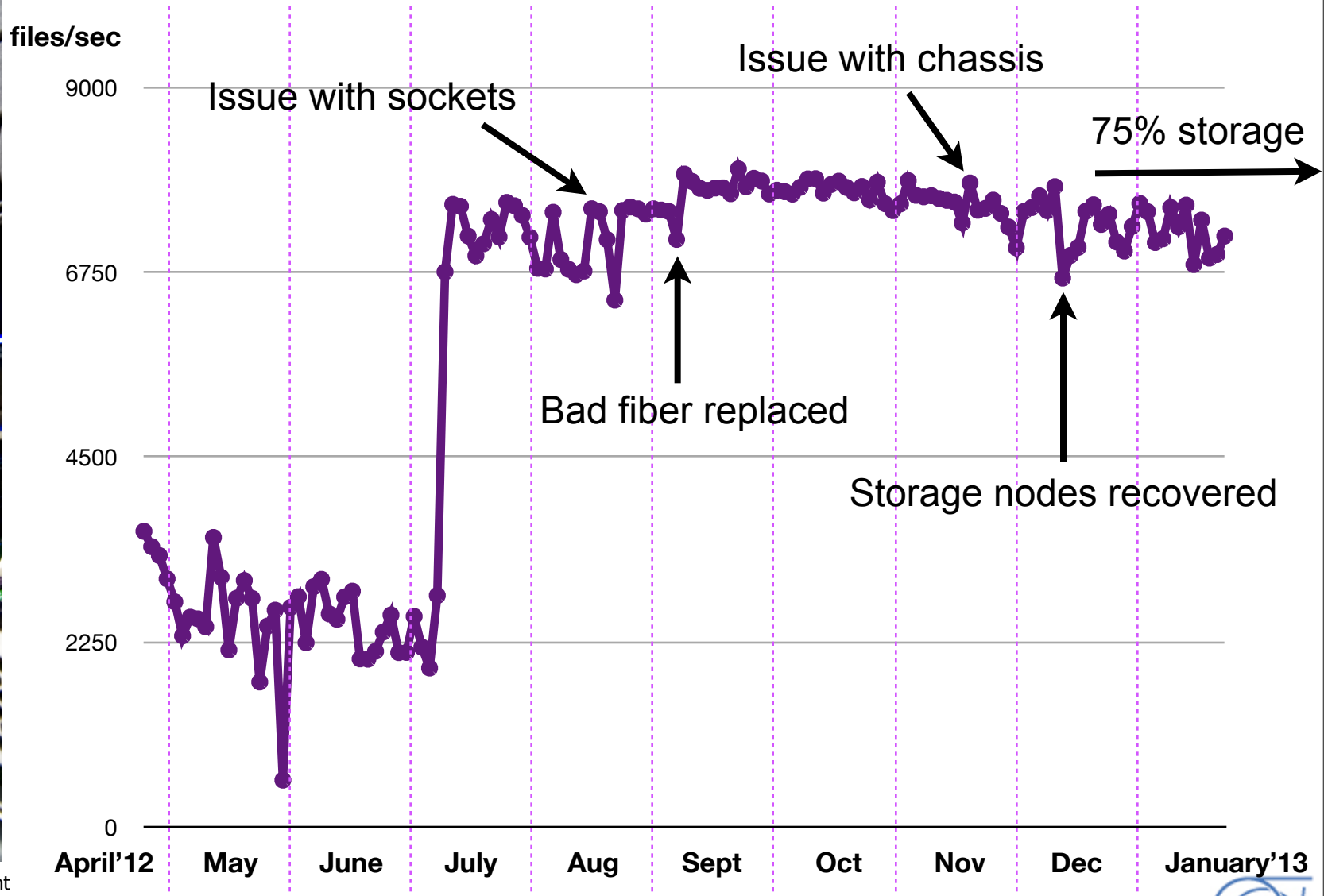
Chassis recovery: Conclusions

- Test has been a success, due to the transfer of data continued normally even without a chassis properly working.
- The rebalance worked, and the active nodes took care of the data in the nodes down, with no failure in any operation.
- The nodes that were down, need around 7 minutes to come back, and another 5 to be fully operative again.
- Less impact in the download performance, than in the upload performance.

Real life chassis unreachable

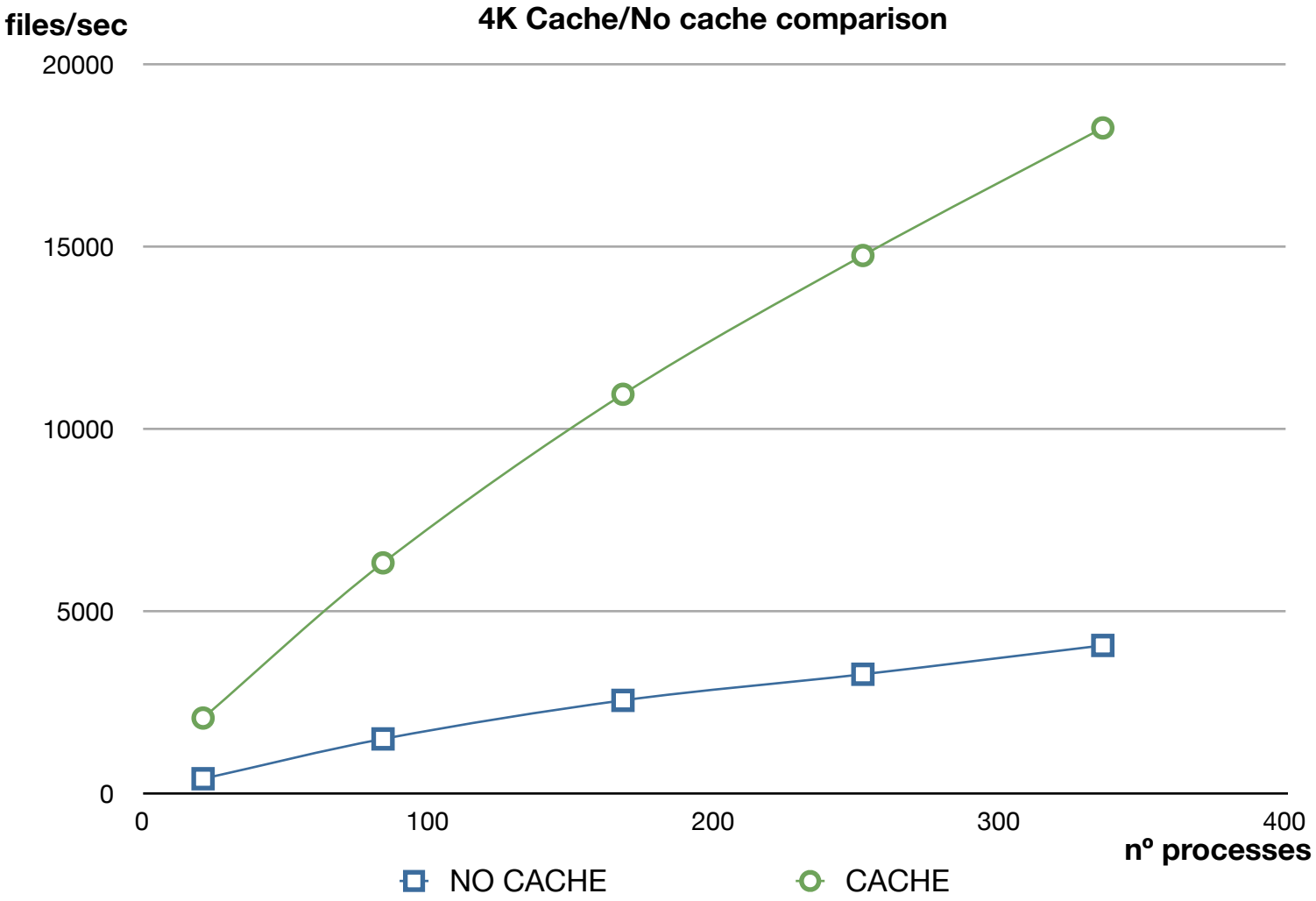
- Oms Portal reports 23 disk failures
  - Among them, chassis 13 completely down
  - Hardware in CC reports only 4 disk failures
- Performance issues in both ups and downs
- Issue recreated in China:
  - Belongs to the BMC system of ARM
- Nodes restored
  - Read performance back, but write still slow
- OMS Portal reports 75% system utilization
  - Measurement of utiization is not correct
  - Lots of data marked as deleted but has not been garbage collected (periodic)

Long term stability test



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

No cache downloads



Works as expected. Metadata download without cache scales linear until 4100 files/sec

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

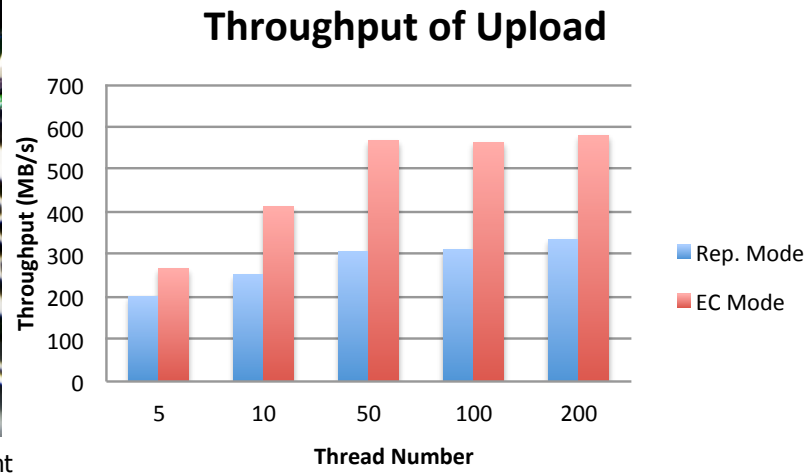
IHEP: UDS

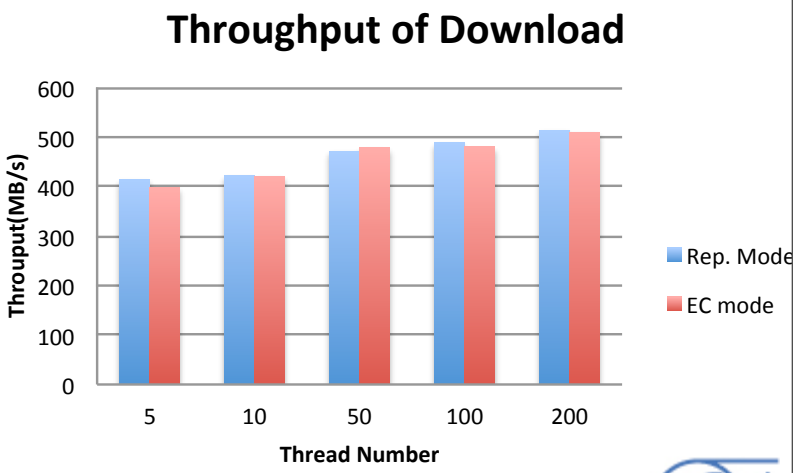
- Another UDS system has been deployed at IHEP, Beijing since October,2012.

- At a smaller scale:
  - 1 mgr node (MN)
  - 2 Service Controller Nodes (OSC)
  - 6 chassis , 96 2TB SoD
- With a newer version
  - V100R001C00SPC101 (IHEP)
  - V100R001C00SPC805T(CERN)
  - Including two new features:
    - “Erasure code”
    - “Multi Data Center”

IHEP: General Performance

- Throughput Test was done with 5 client boxes, 5 accounts, 500 buckets, 100MB objects in both replication mode (replication=3) and EC mode(9 data, 3 parity).
- Limited by the physical bandwidth of client boxes (5Gb/s), except Rep. download test, all the other three tests have not reached the theoretical peak performance of the system.





Conclusion

- Tested OSC scalability performance
  - Work distributed properly between all boxes
- Tested metadata performance
  - Expected scalability in writes when enough buckets provided
  - 18000 files/second in reads
- Tested total throughput up to 18 Gb/s
  - fully maxed out the 2 fibres available
  - balanced system with 350MB/s per OSC
- Recovery after powering off a chassis
  - Transparent disk failure recovery proved
- Technical problems found and solved

Future plans

- Short term
  - Keep analysing performance impact of read cache
  - Analyse performance impact of write cache and journal
  - 21 client transparent disk failure recovery test
  - Excercise transparent upgrade procedure
  - New fellow starting 1st of March
- Long term
  - Second PB system with consumer disks arrive May 2013
  - Upgrade system versions at CERN and IHEP
  - Joint test with IHEP to verify the remote datacenter feature
  - Erasure code impact on performance and space overhead
  - Evaluation with jobs from experiments
  - Prove TCO gains of the system as part of a production service

DSS

Data & Storage Services

CERN IT Department

Huawei Cloud Storage

Maitane Zotes Resines, CERN IT
(With contributions from Wang Lu, IHEP Beijing)

Openlab Major Review Meeting
31. January 2013
CERN, Geneva

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

HUAWEI

CERN openlab

IHEP: Issues

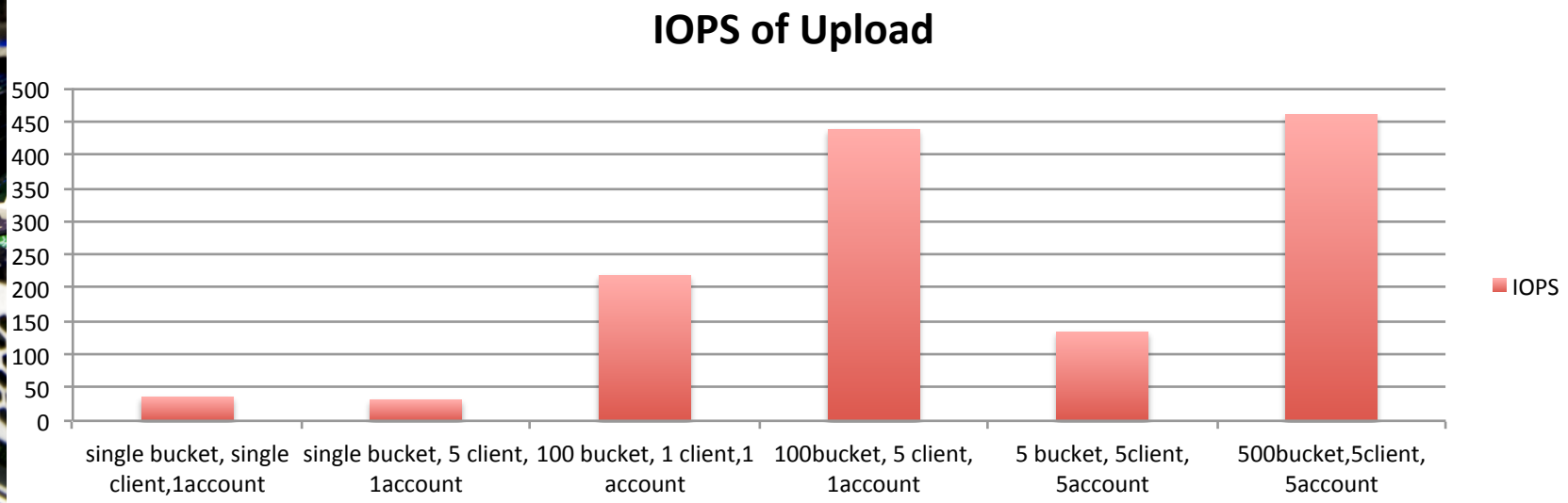
- Metadata inconsistencies
  - Listing the contents of reportedly existing buckets gives ‘bucket does not exist’ error
  - ‘internal error’ retrieved (even after several retries) when listing bucket contents
  - Successfully deleted objects may “reappear” when the contents of the bucket are retrieved
  - Huawei has provided a patch for the first two problems. To verify the patch, further tests are still needed. The last problem cannot be regenerated, detailed trace information provided to Huawei
- Management interface
  - Impossible to retrieve the list of existing users in the system: need of a separate database to keep this information
  - Cannot create more than 2 pair of credentials (access key, secret key) per user: impossible to create “volatile” credentials
  - Cannot reuse a recently deleted userid before an expiration period (not documented)
  - email addresses of users need to be unique, so cannot have more than one pair of credentials per individual, unless using fake email addresses

Improvements to ROOT

- Modification of ROOT to improve read-only file access capabilities using S3 protocol
  - Accepted by ROOT development team
- Included improvements:
  - Support of both HTTP and HTTPS as transport protocols
  - Provider-agnostic: previous implementation basically worked only against Amazon S3 servers
  - Can exploit multi-range requests if the server supports them: Huawei does, Amazon does not
  - Support of authentication information on a per-file basis: useful when dealing with files hosted by different providers

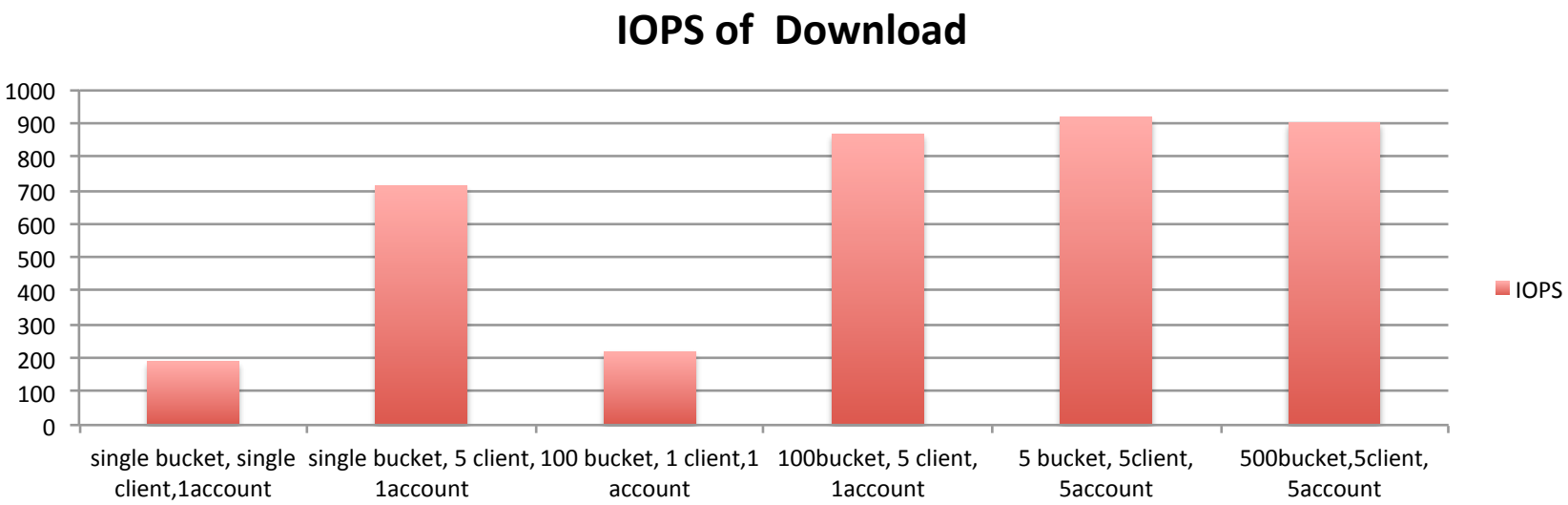
IHEP: General Performance

- With 5 client boxes, 4KB objects, IOPS test of upload in replication mode(replication=3). With this configuration, peak performance of the system is 450 files/sec.



IHEP: General Performance

- With 5 client boxes, 4KB objects, IOPS test for Download in replication mode(replication=3). With this configuration, peak performance of the system is 900 files/sec.



OSC scalability

- **Aim:** Check the proper distribution of the data between all the OSC boxes.
- Each OSC should deal with the same amount of data

Test 1	1 OSC	3 Clients	30 Processes
Test 2	2 OSC	6 Clients	60 Processes
Test 3	3 OSC	9 Clients	90 Processes
Test 4	4 OSC	12 Clients	120 Processes
Test 5	5 OSC	15 Clients	150 Processes
Test 6	6 OSC	18 Clients	180 Processes
Test 7	7 OSC	21 Clients	210 Processes