Huawei Cloud Storage

Maitane Zotes Resines, CERN IT

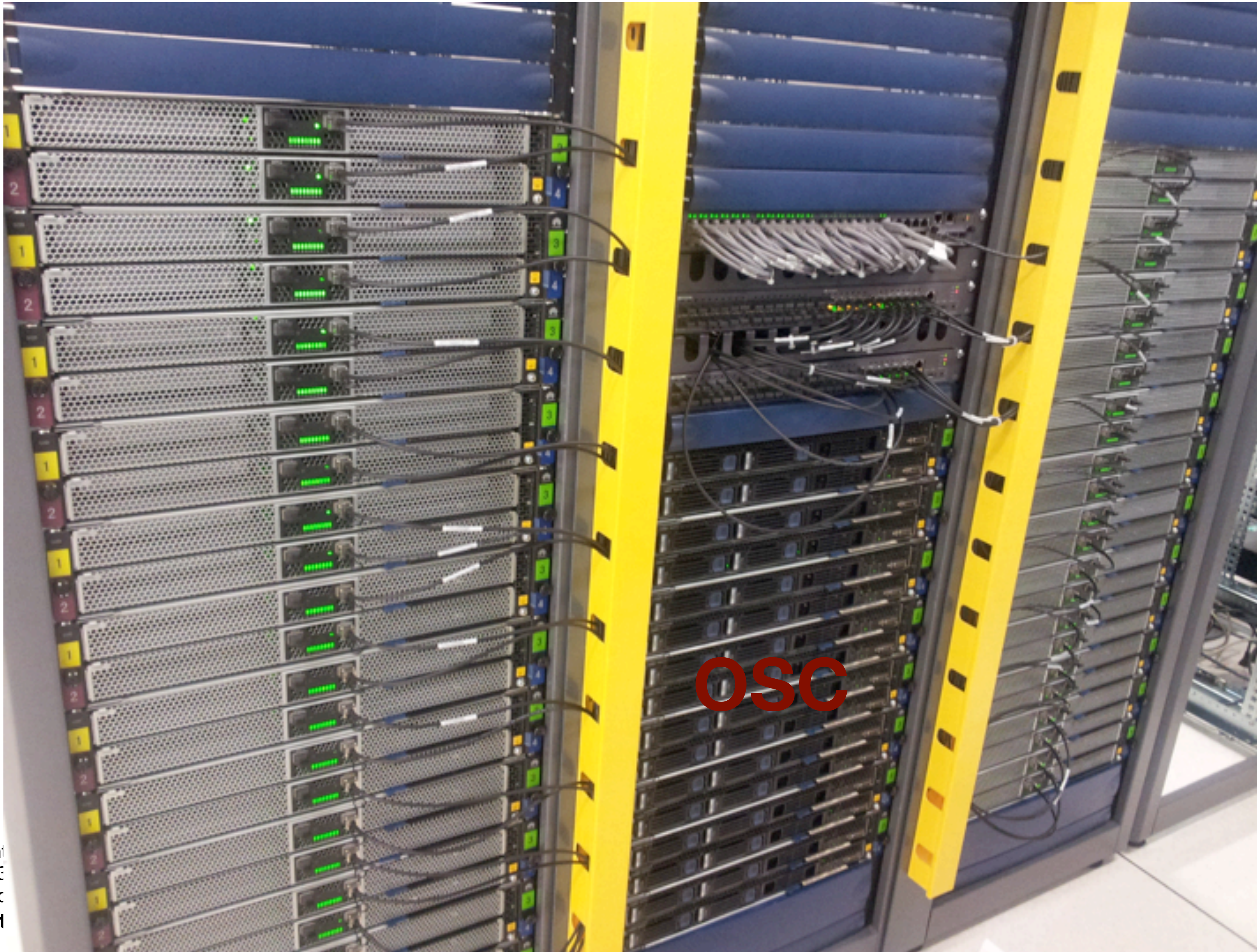
Openlab Major Review Meeting
27. September 2012
CERN, Geneva

Outline

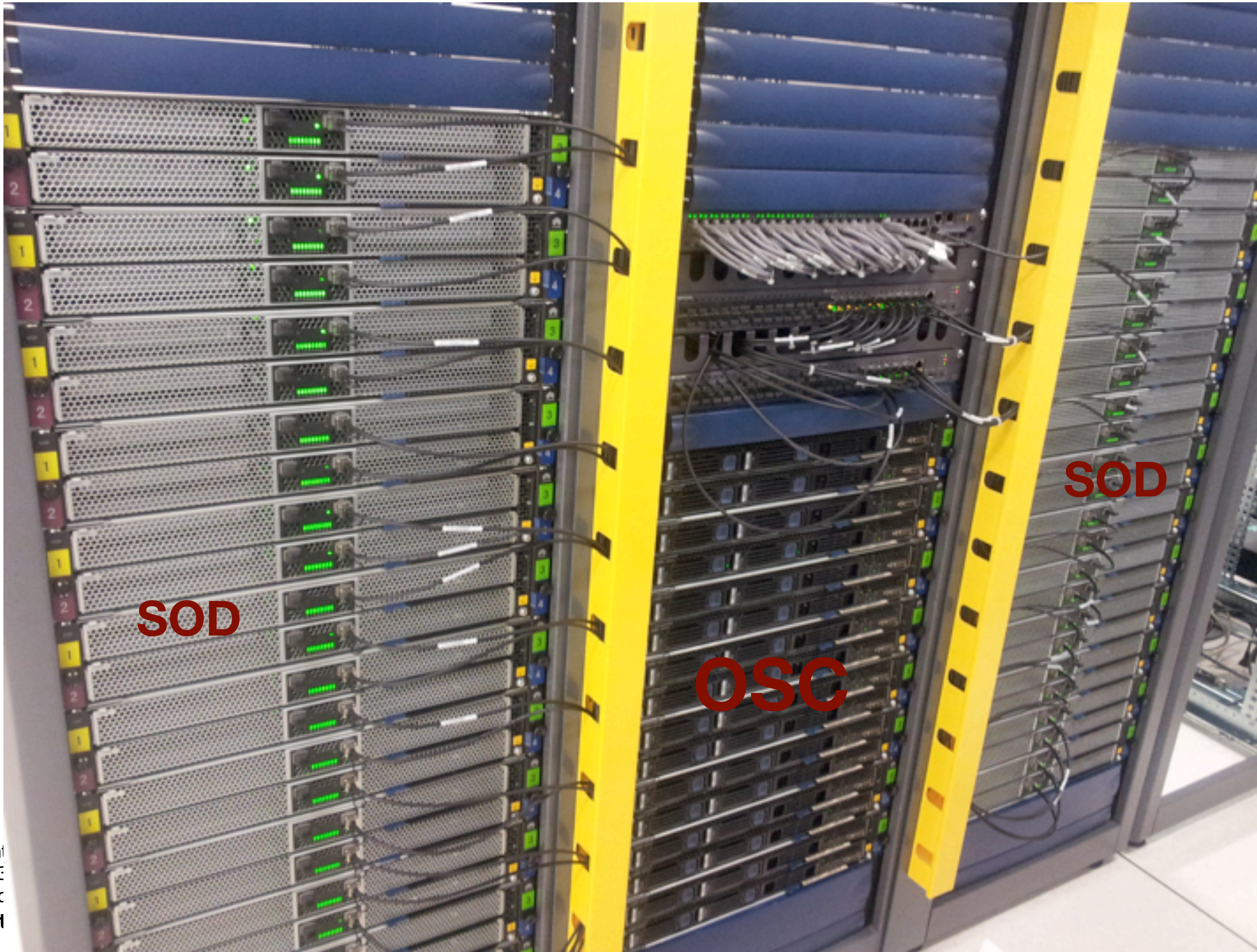
- Huawei setup
- Remaining Issues after the first phase
  - see last major review
- Benchmark execution framework
- Initial results
- S3 access with ROOT and S3FS
- System upgrade
- New results and comparison
- Future workplan

Huawei Setup

Huawei Setup

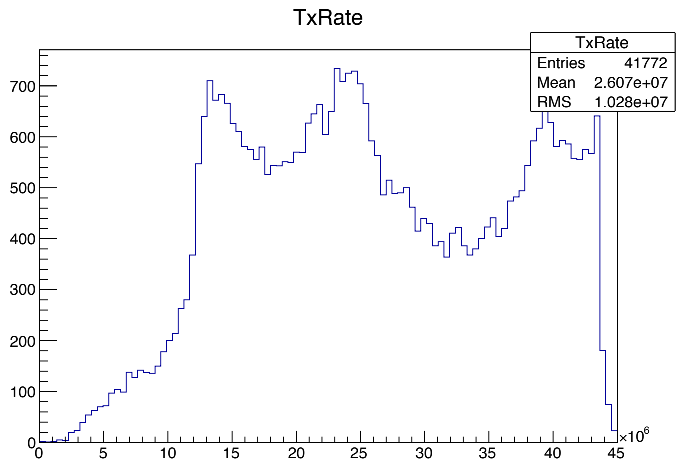


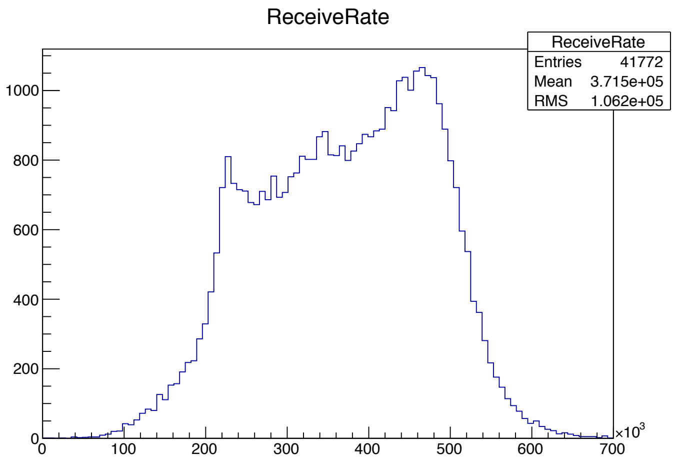
Huawei Setup



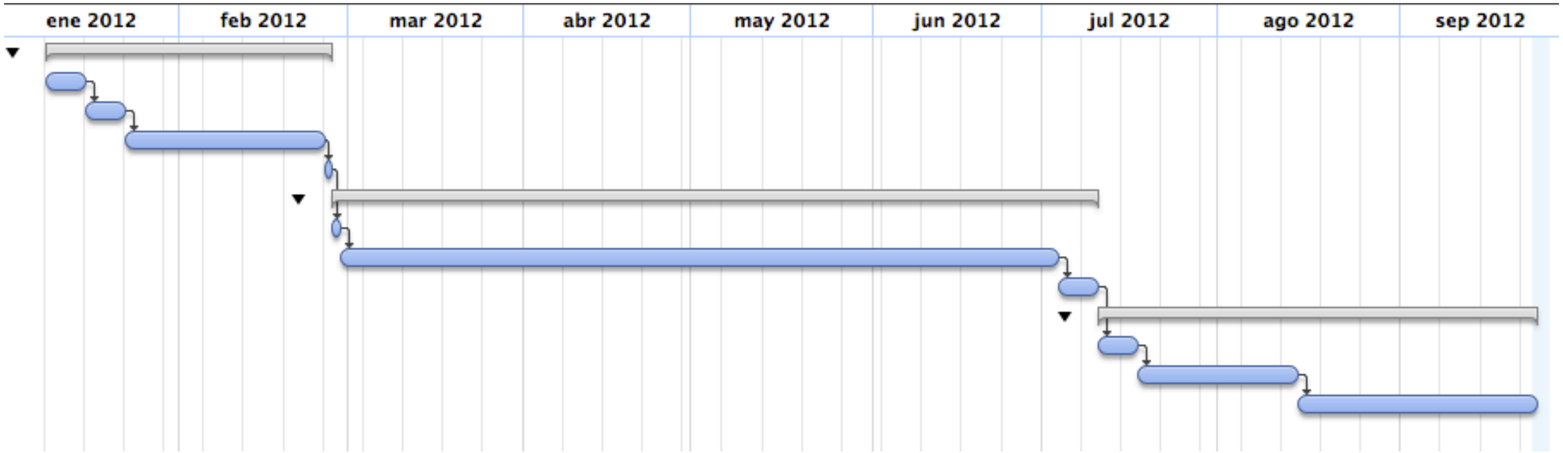
Benchmark execution framework

- New framework integrated with ROOT
  - Histograms
  - Integrated with the Python benchmark
  - ssh connection to clients using lxplus





Timeline



1st Phase

Project start
Mounting of system
First tests
First Review

2nd Phase

Obtain 5 client boxes
Redo tests with new clients
Upgrade of system

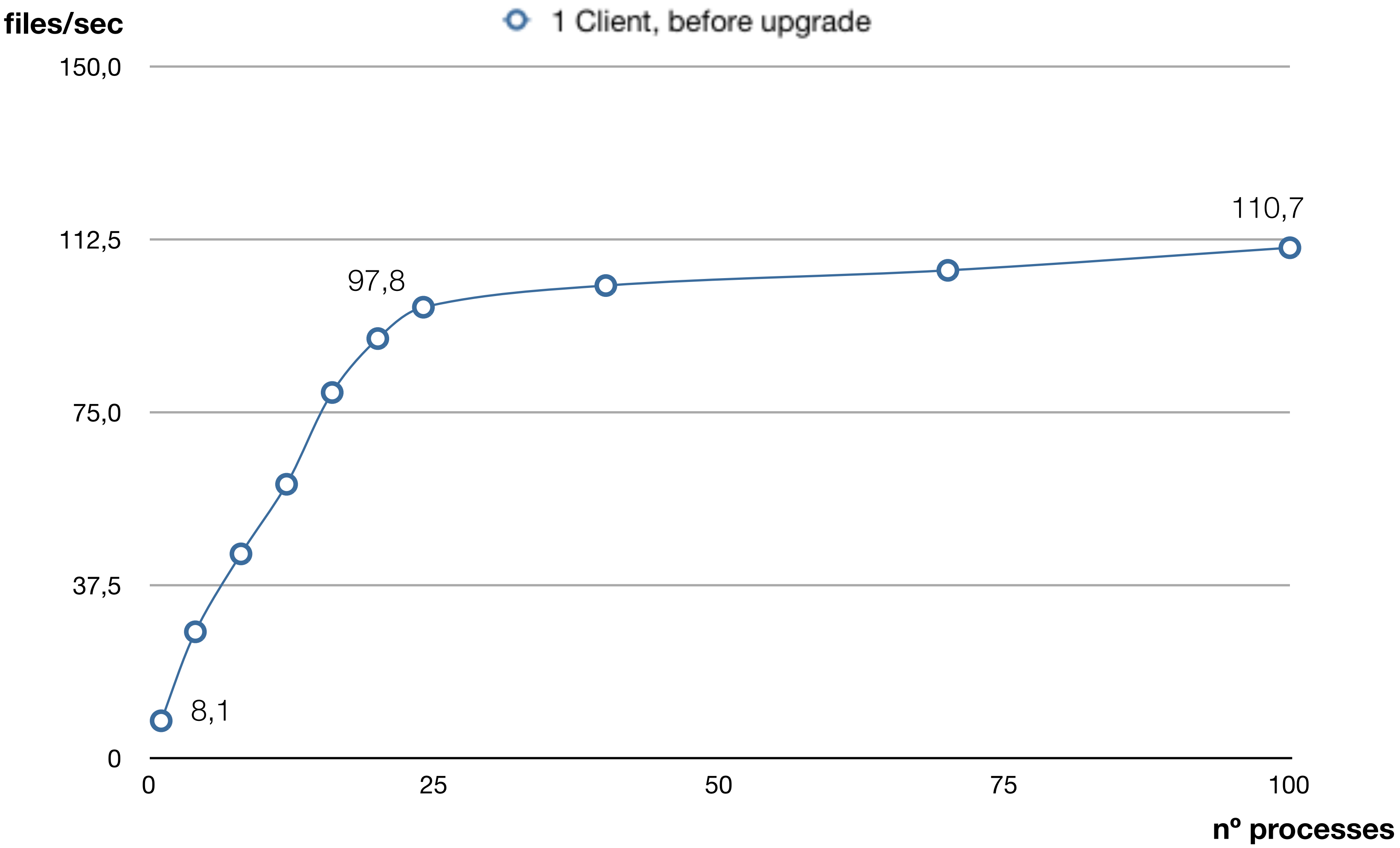
3rd Phase

Redo tests
Download issue found
Fix issue
Redo download tests
Obtain 20 client boxes
Start 20 client tests
Openlab Major Review

Limitations observed in first testing phase

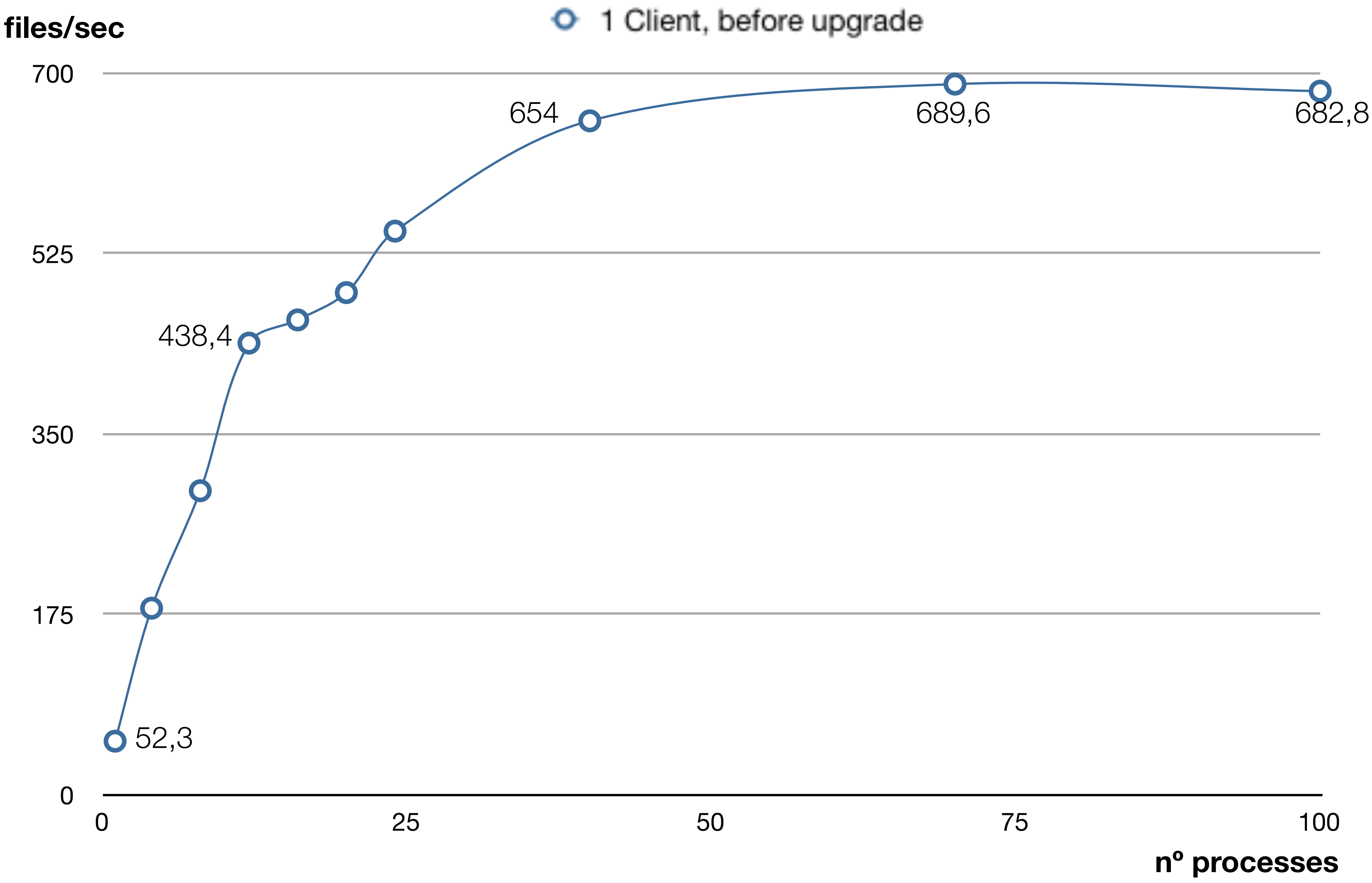
- Upload limit
  - 50 files/second
- Download limit
  - 425 files/second
- Benchmark issues
  - Delayed closing of client sockets produces socket shortage
  - Client bandwidth limit
  - Client memory limit

4KB Upload - Metadata Scaling



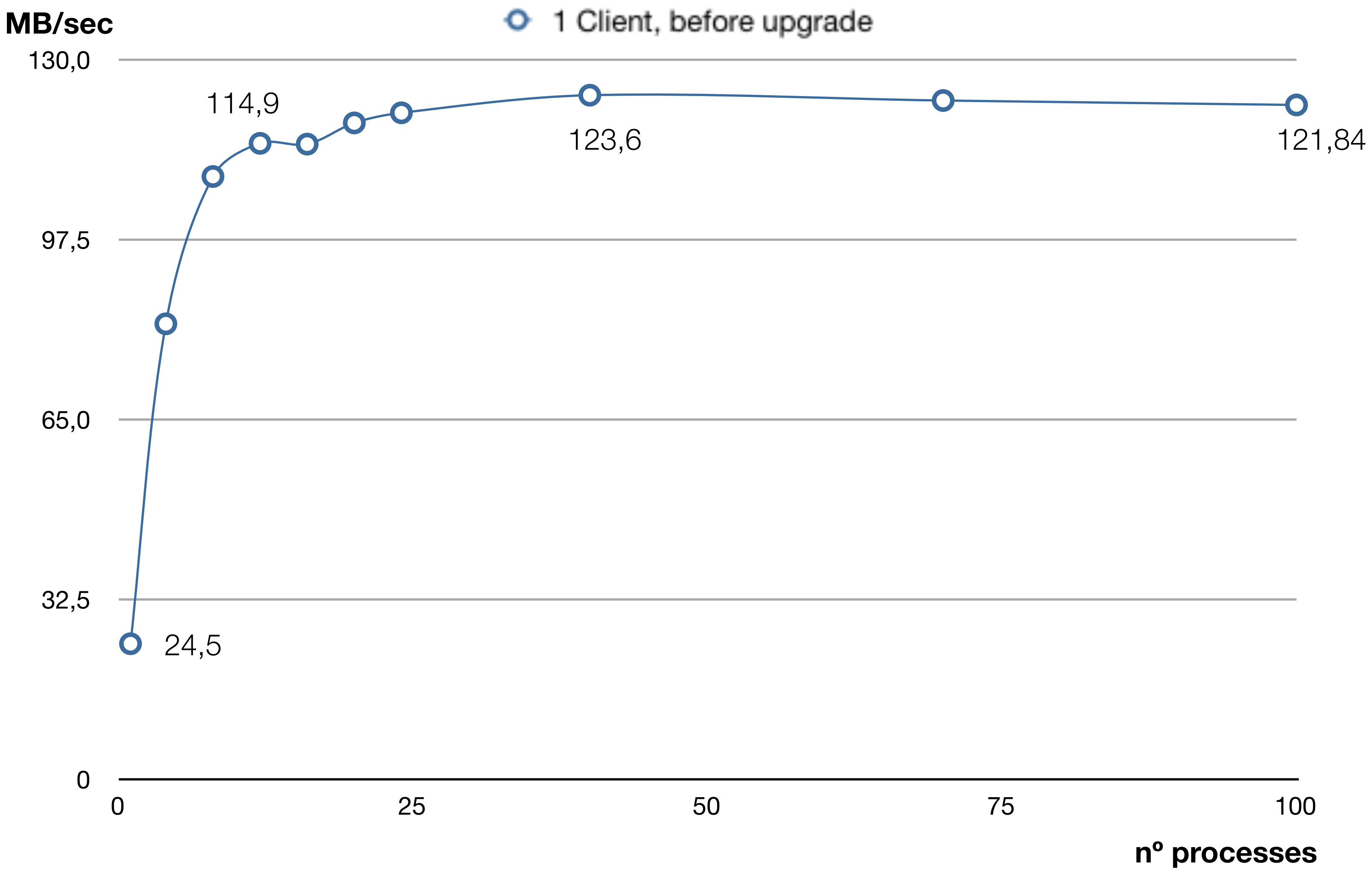
Succesfull scalability of metadata

4KB Download - Small Reads



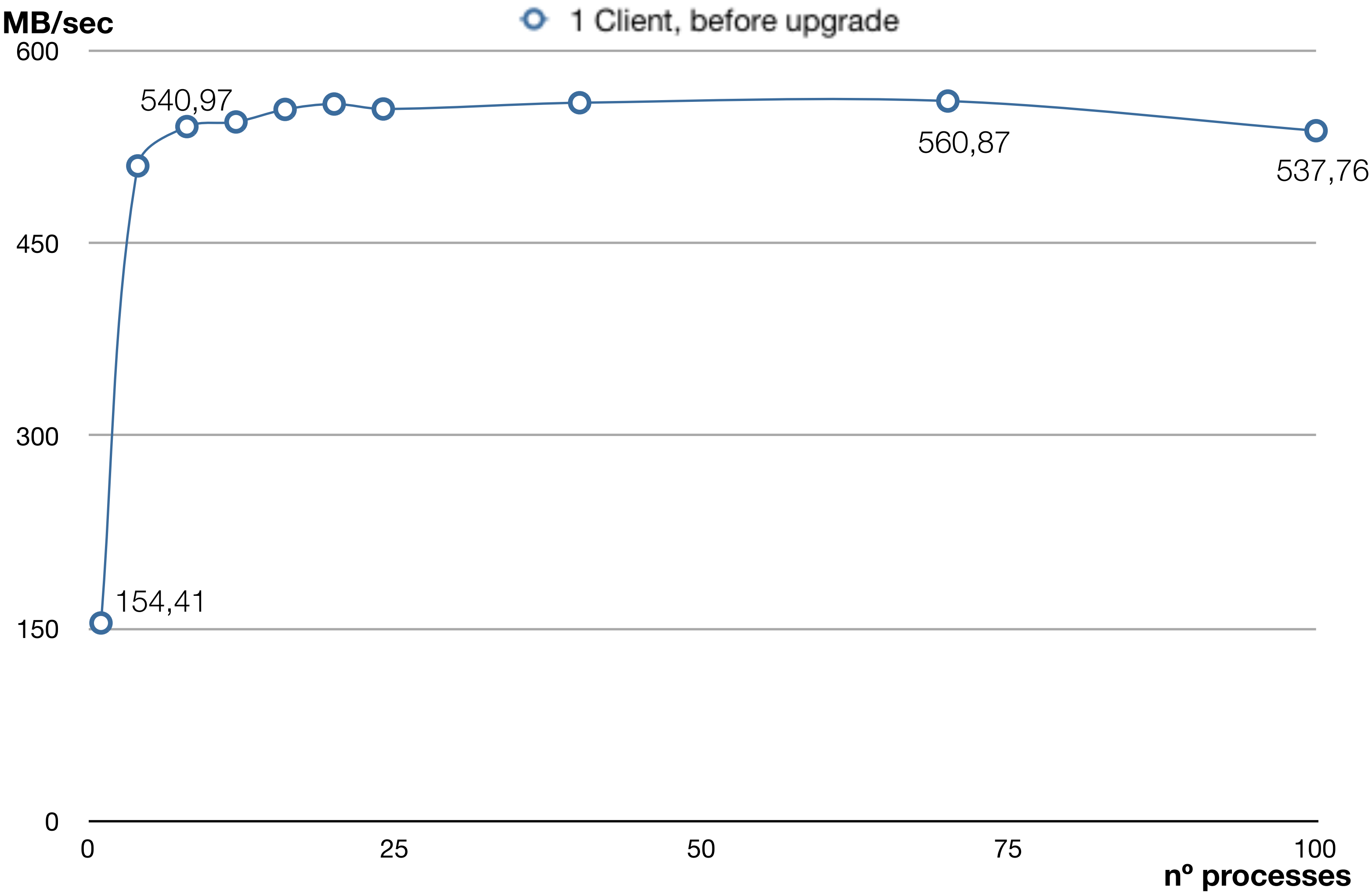
Successful scalability of metadata

100MB Upload - Throughput Scaling



Scales properly and bandwith limit of 5Gb reached

100MB Download - Sequential access of bulk data



Scales properly and bandwith limit of 5Gb reached

Observations

- The system behaves as expected
- 4KB Tests
  - 110,7 files/sec upload limit
  - 682,8 files/sec download limit
- 100MB Tests
  - Reached the client bandwidth limit
  - Stable

Integration with ROOT

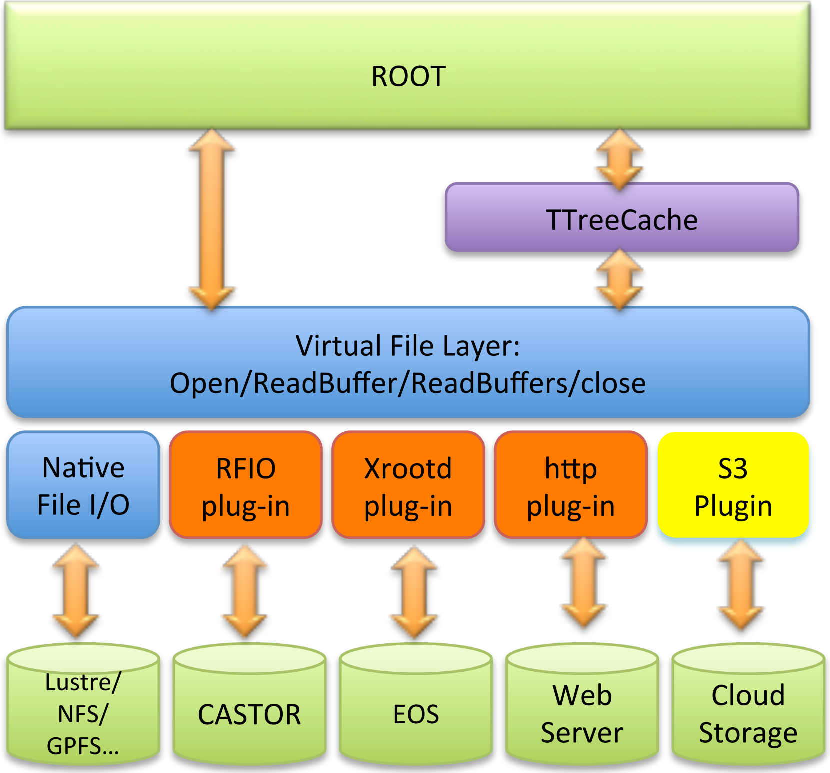
- Huawei UDS is a typical cloud storage which provides Amazon S3 Interface

- ROOT is designed to support different I/O protocol
  - rfio, xrootd, posix …

- Two solutions to run ROOT on cloud storage
  - develop a S3 Plug-in for ROOT (validated by tests)
  - develop a fuse-based POSIX file system to mount cloud storage

S3 plug-in for ROOT

- 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 API
- Integrated with the distributed I/O test framework

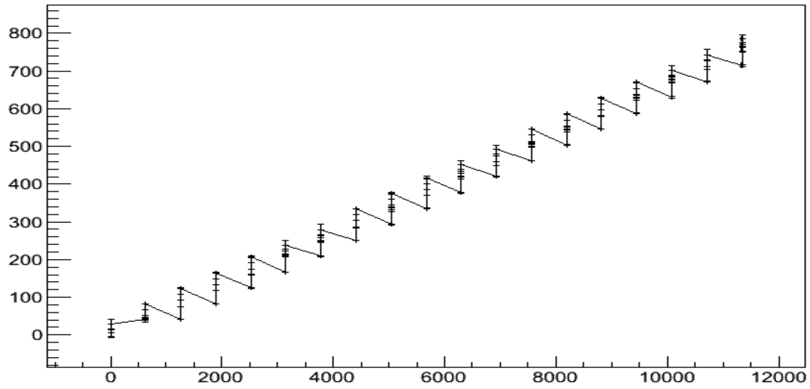


Fuse-based file system

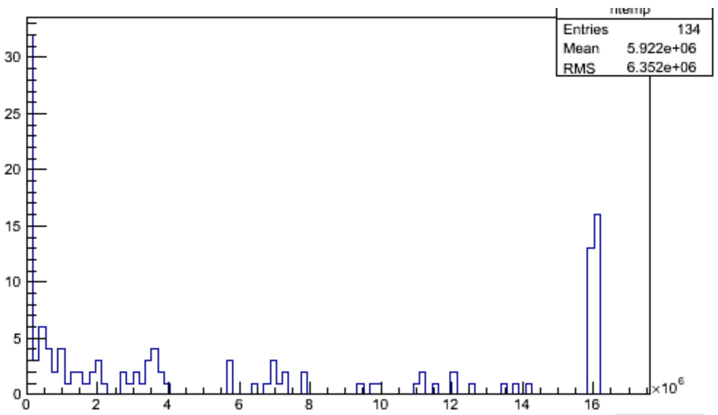
- Open source prototype S3FS
  - http://code.google.com/p/s3fs/
- Current limitations:
  - Can only mount one single bucket instead of the whole system
  - Instead of remote I/O, file is downloaded to local cache during “open”
  - “df” returns not relevant information

Test Job

- 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*”



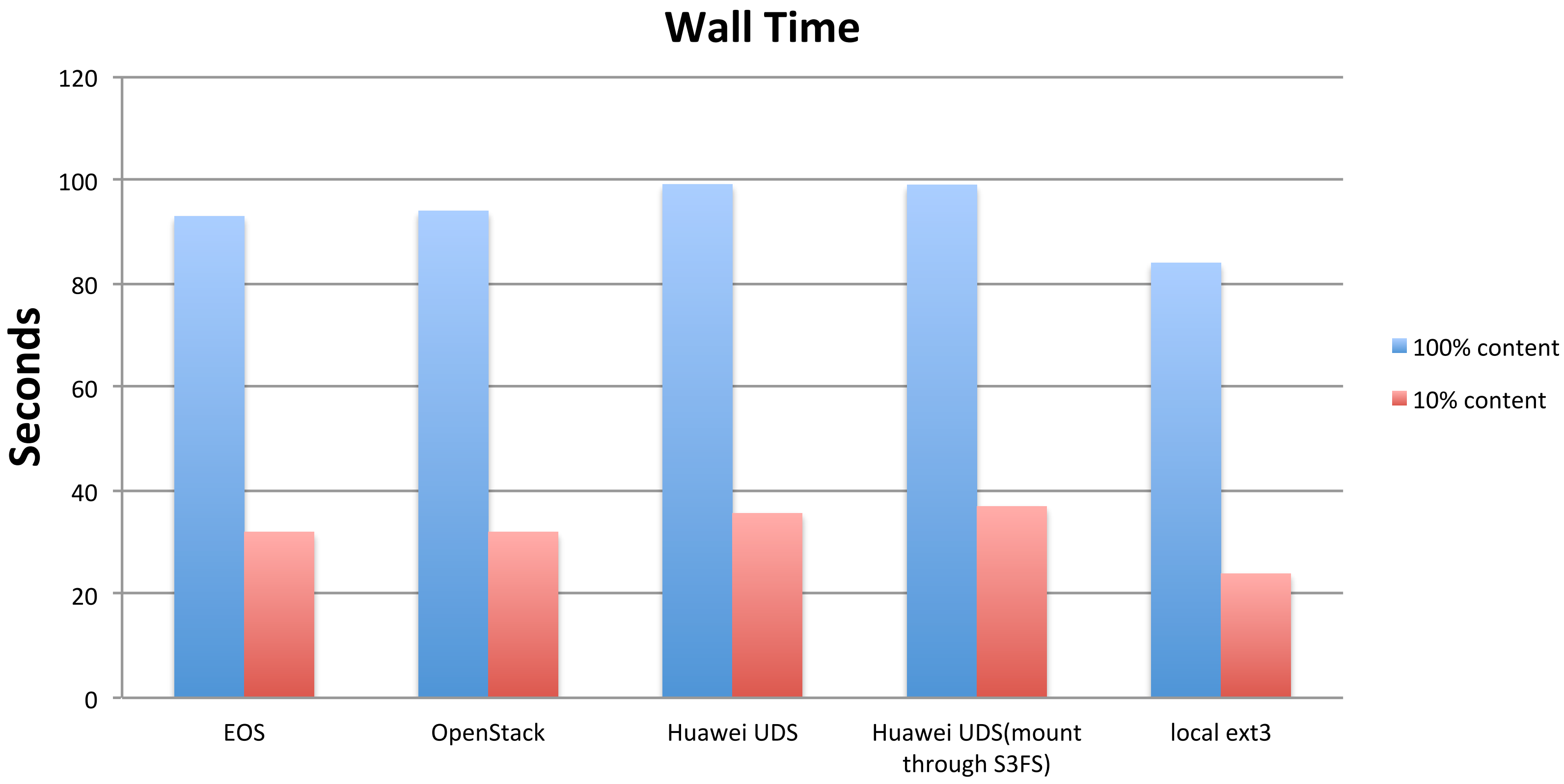
Reading Offset vs Entry



Distribution of read Size

Test Result

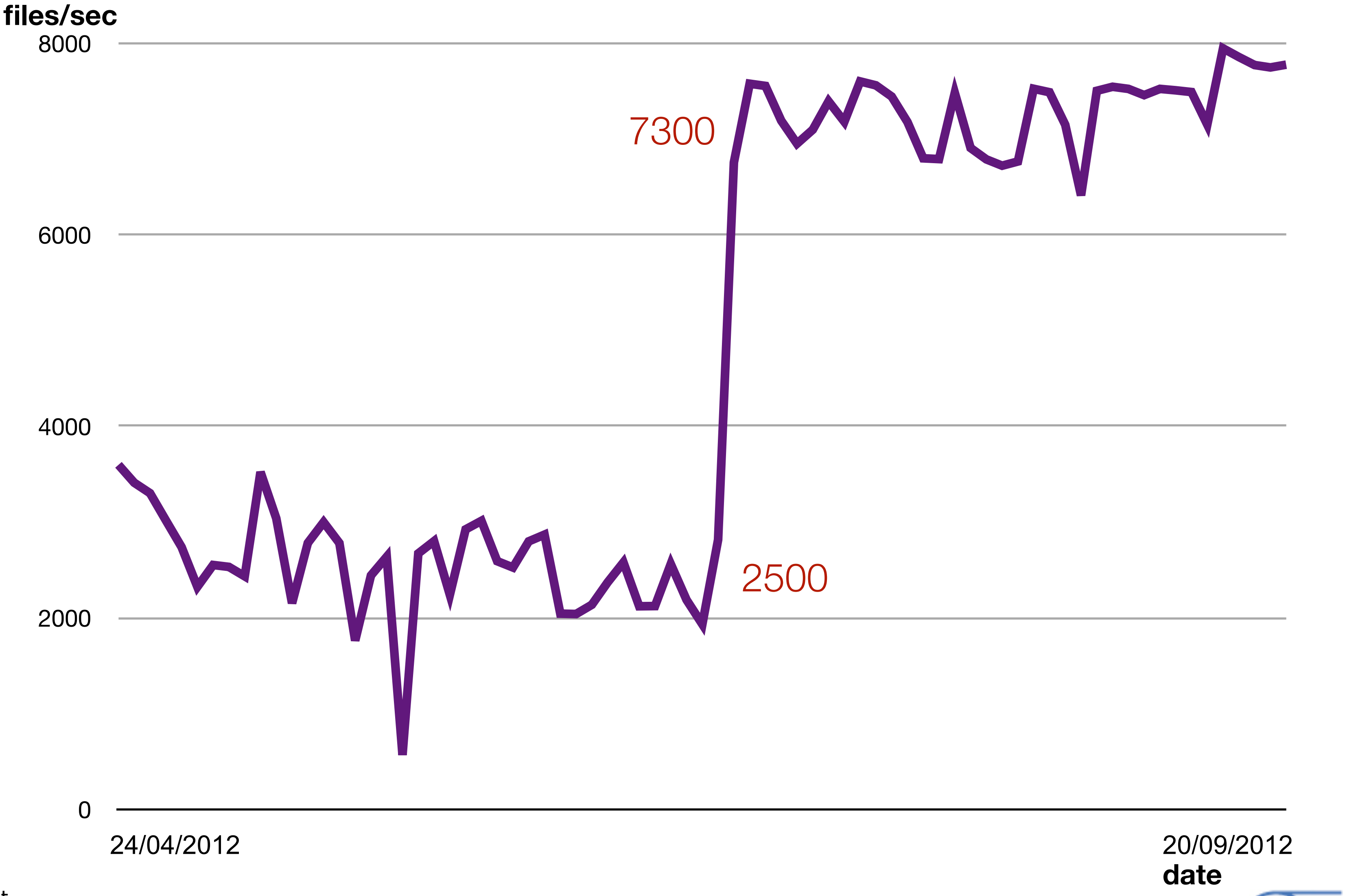
- Single client



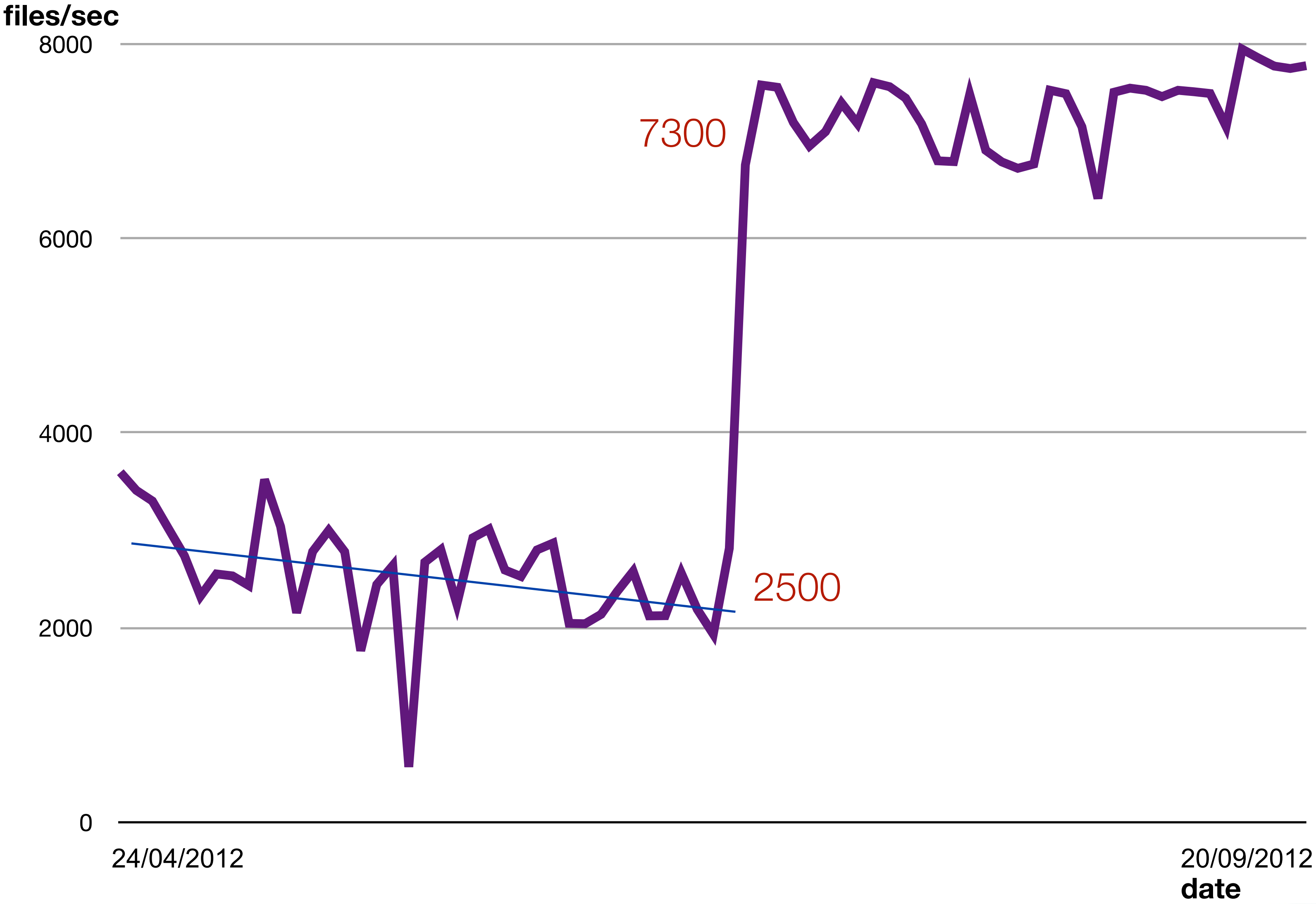
Upgrade of system software

- Hardware improvements
- Upgrade of SOD node
- Performance optimization
- Billing function
- User management API
- More S3 features
- Cache turned on/off

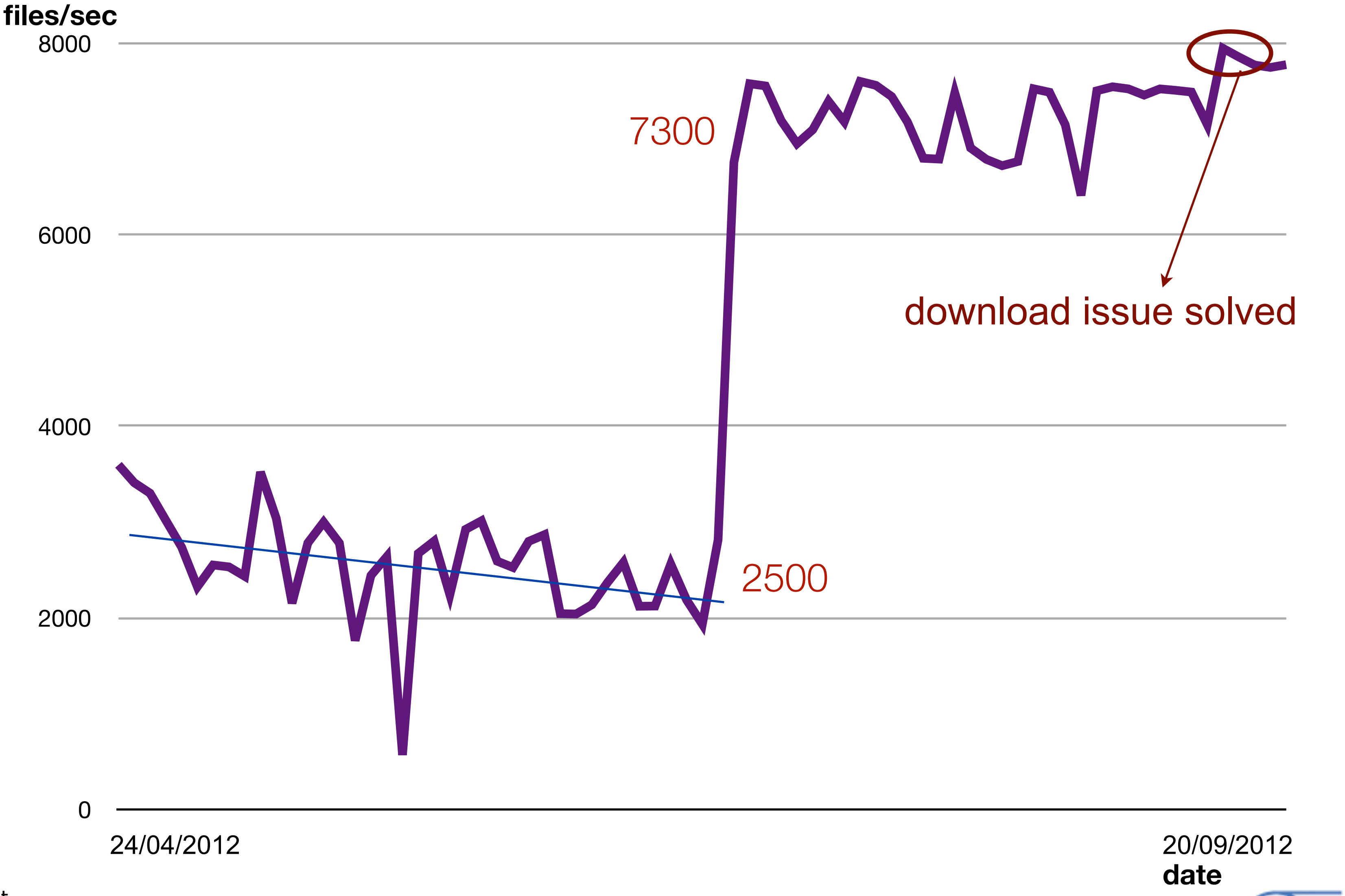
Morning test - Long term stability



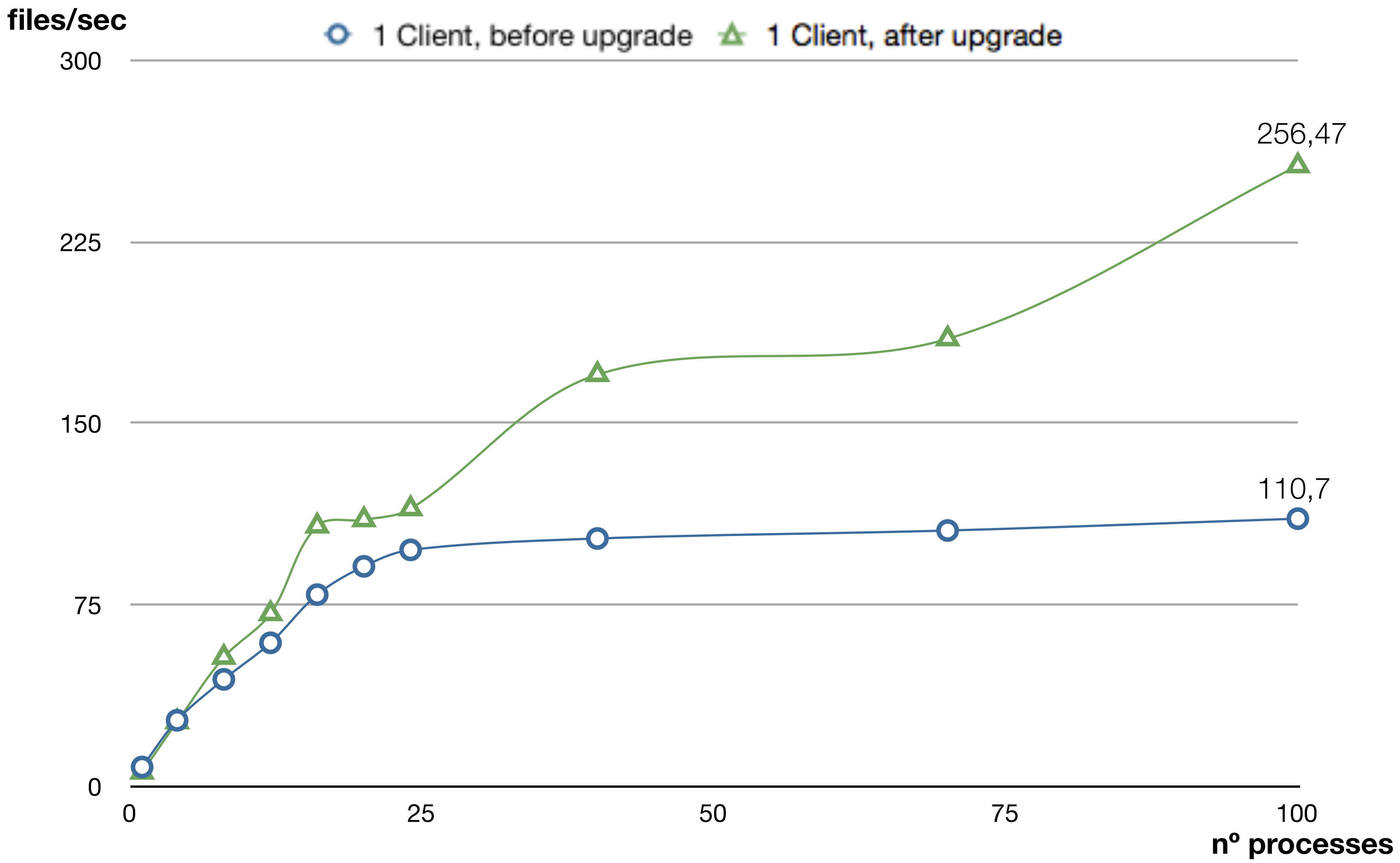
Morning test - Long term stability



Morning test - Long term stability



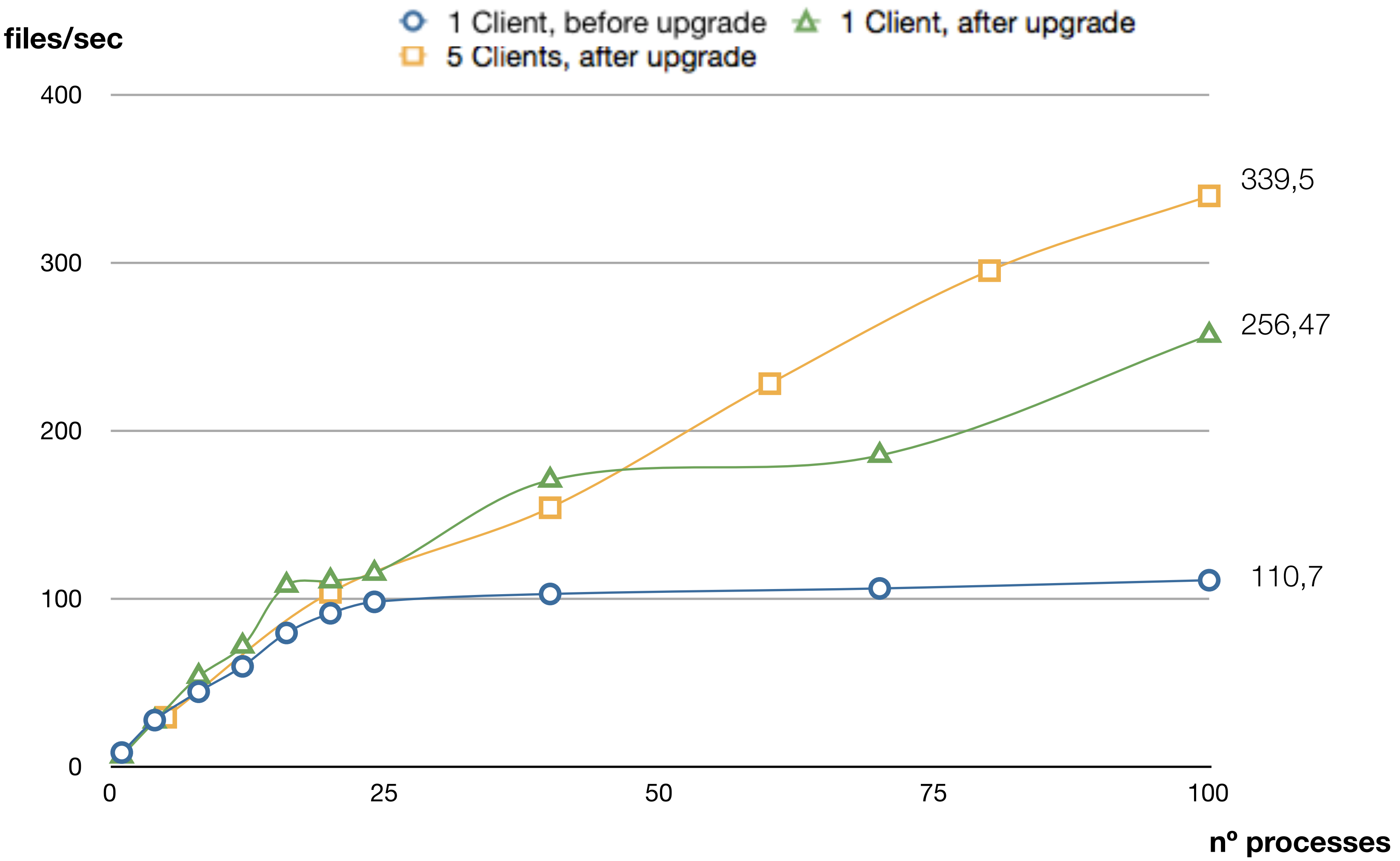
4KB Upload



After the upgrade the performance doubles

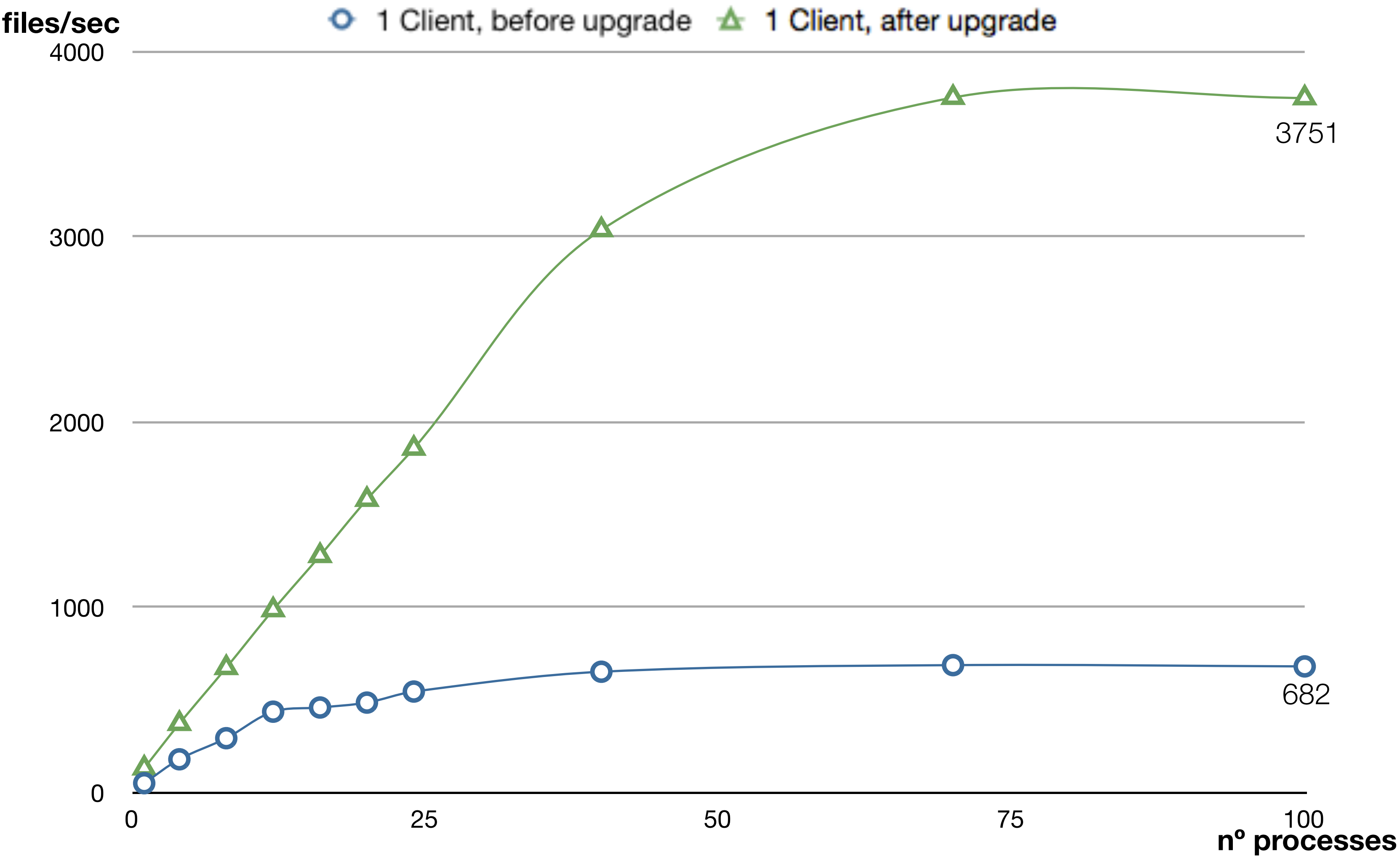
4KB Upload

382,45


files/sec
○ 1 Client, before upgrade
△ 1 Client, after upgrade
□ 5 Clients, after upgrade
400
300
200
100
0
0
25
50
75
100
n° processes
339,5
256,47
110,7

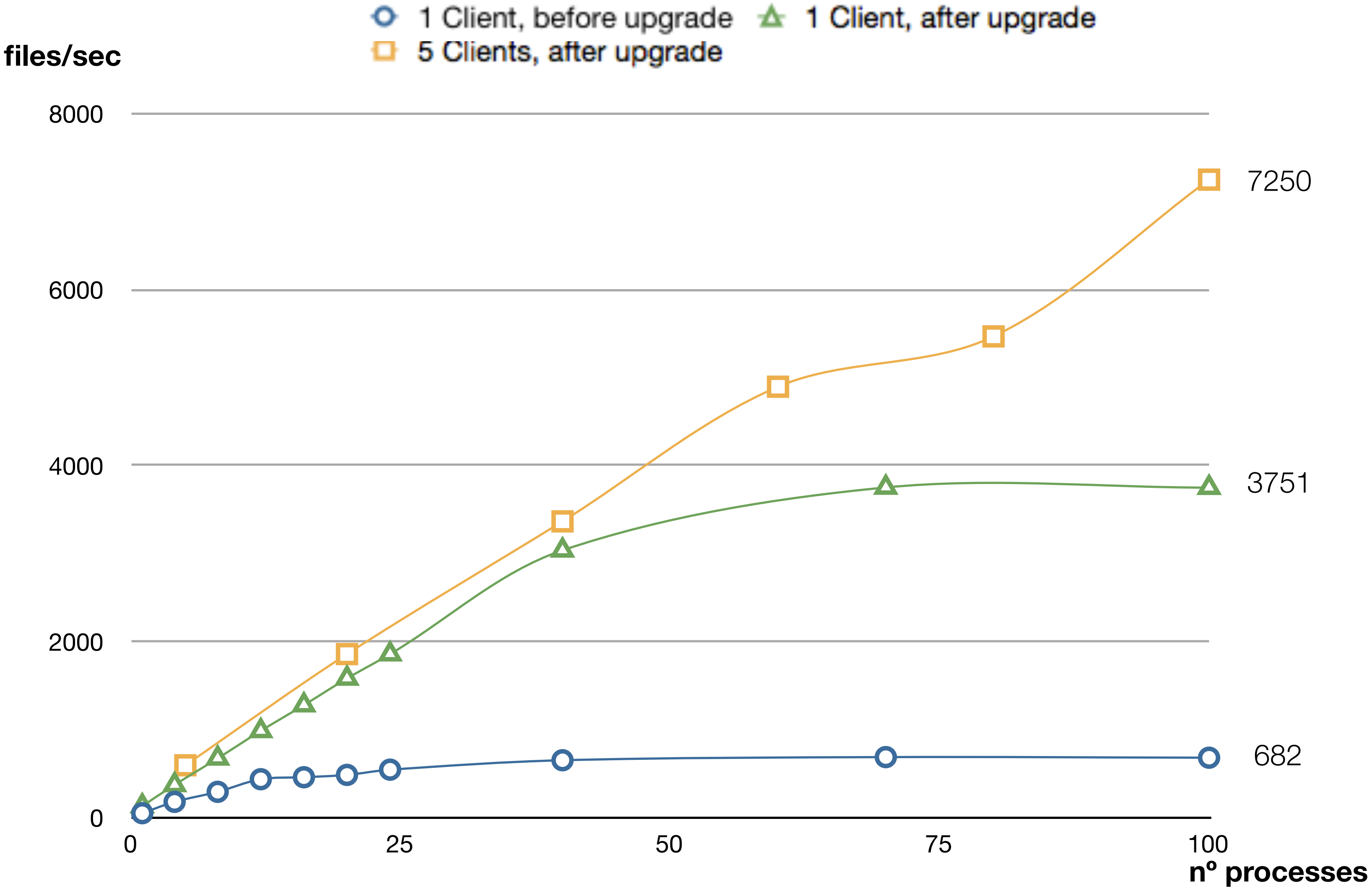
5 client tests improve the numbers

4KB Download



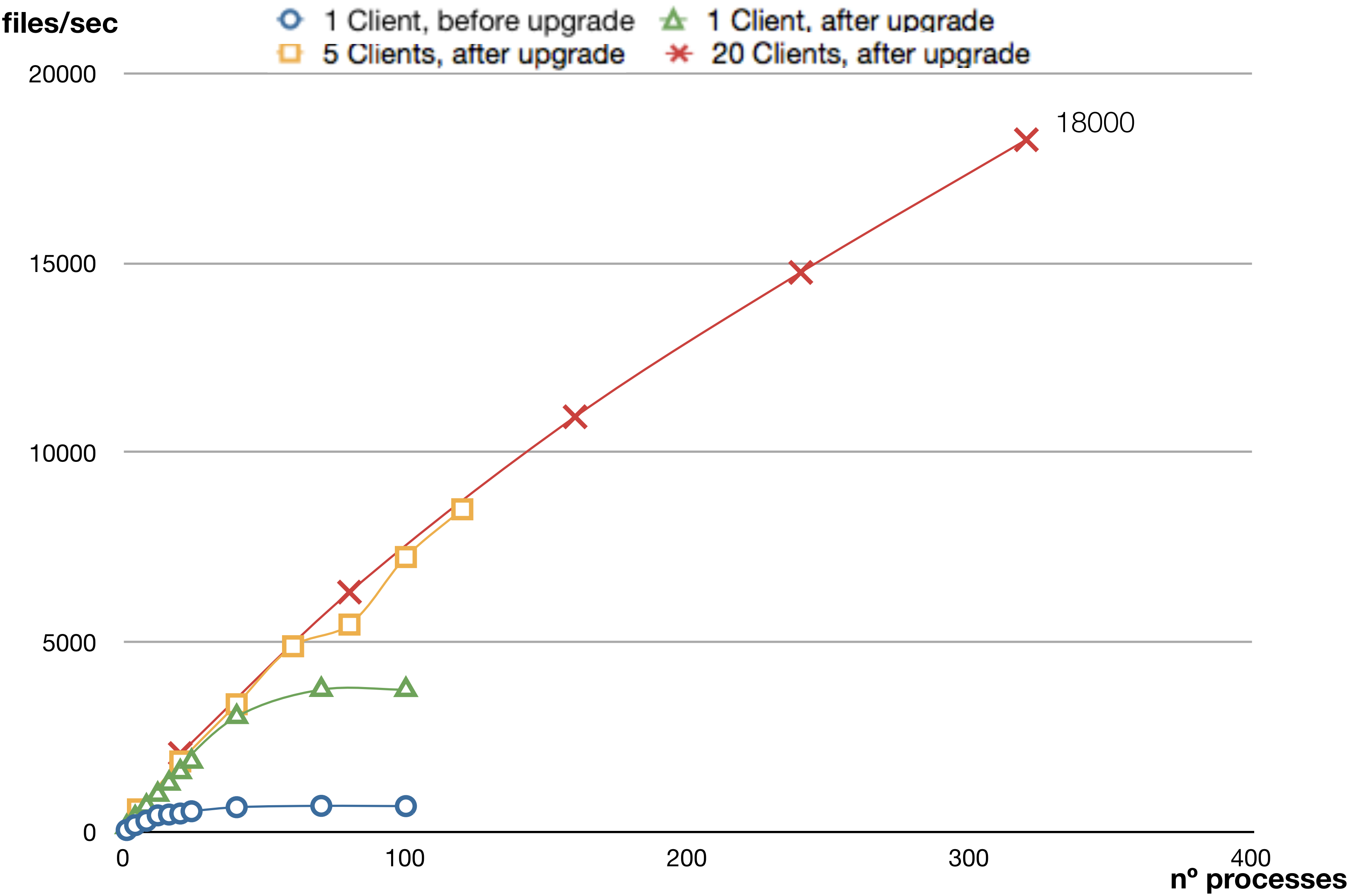
Performance after the upgrade five times better

4KB Download


○ 1 Client, before upgrade
△ 1 Client, after upgrade
□ 5 Clients, after upgrade
files/sec
8000
6000
4000
2000
0
0
25
50
75
100
nº processes
7250
3751
682

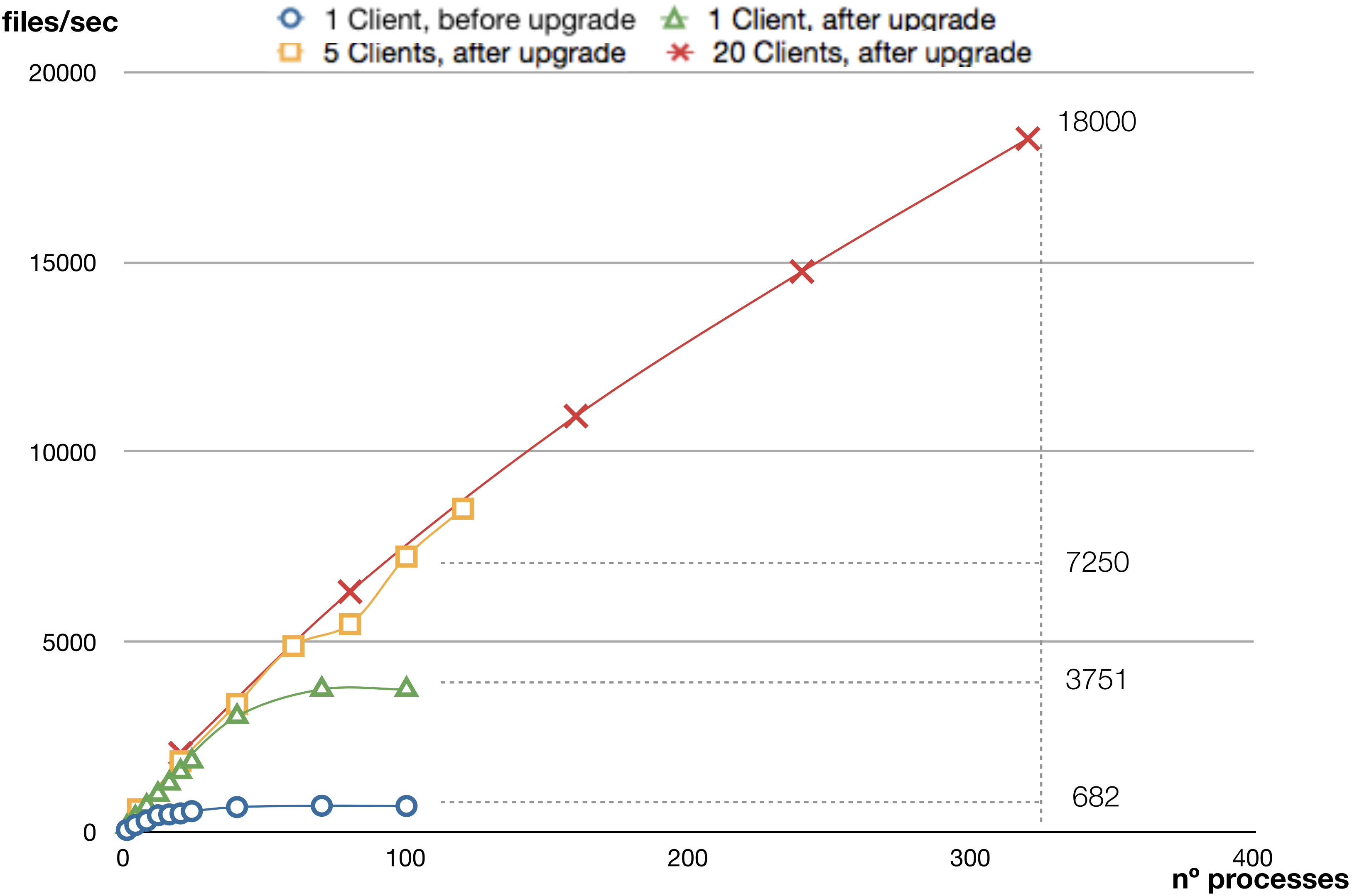
5 client test doubles last better result

4KB Download


files/sec
○ 1 Client, before upgrade
△ 1 Client, after upgrade
□ 5 Clients, after upgrade
✳ 20 Clients, after upgrade
20000
15000
10000
5000
0
18000
0
100
200
300
400
nº processes

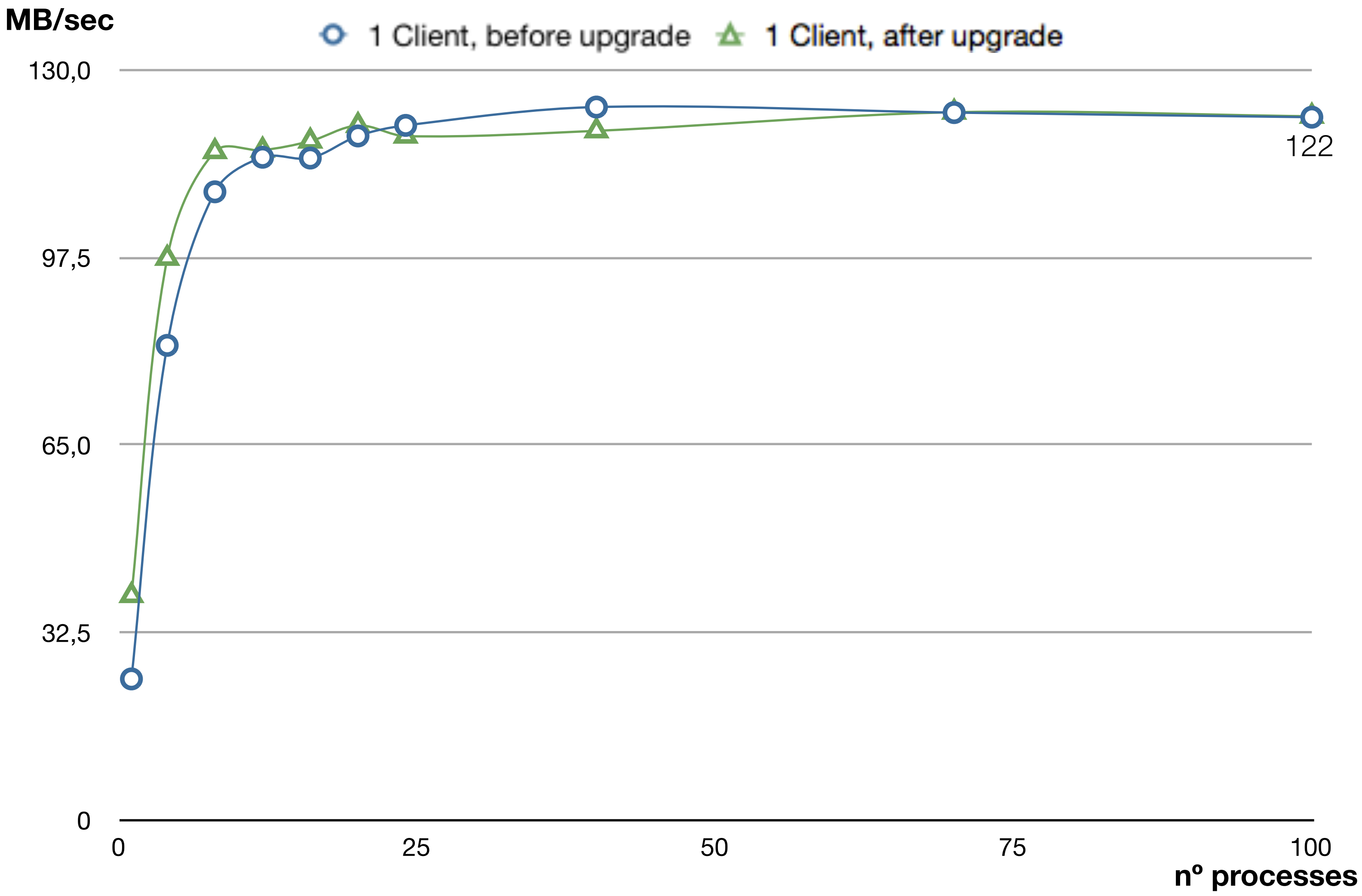
Succesfull scalability of metadata in 20 boxes

4KB Download


files/sec
○ 1 Client, before upgrade
△ 1 Client, after upgrade
□ 5 Clients, after upgrade
✳ 20 Clients, after upgrade
20000
15000
10000
5000
0
18000
7250
3751
682
0
100
200
300
400
nº processes

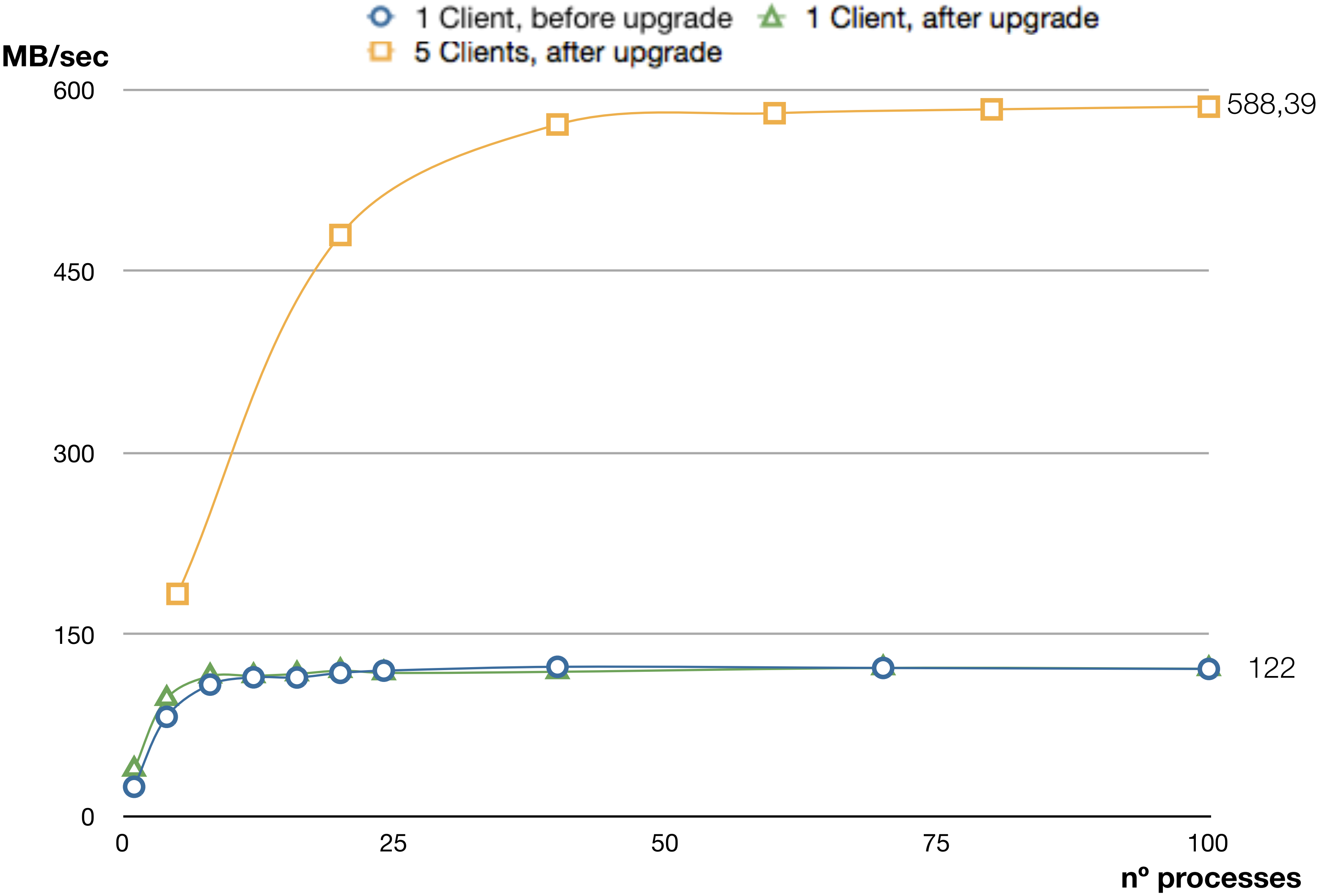
Succesfull scalability of metadata in 20 boxes

100MB Upload



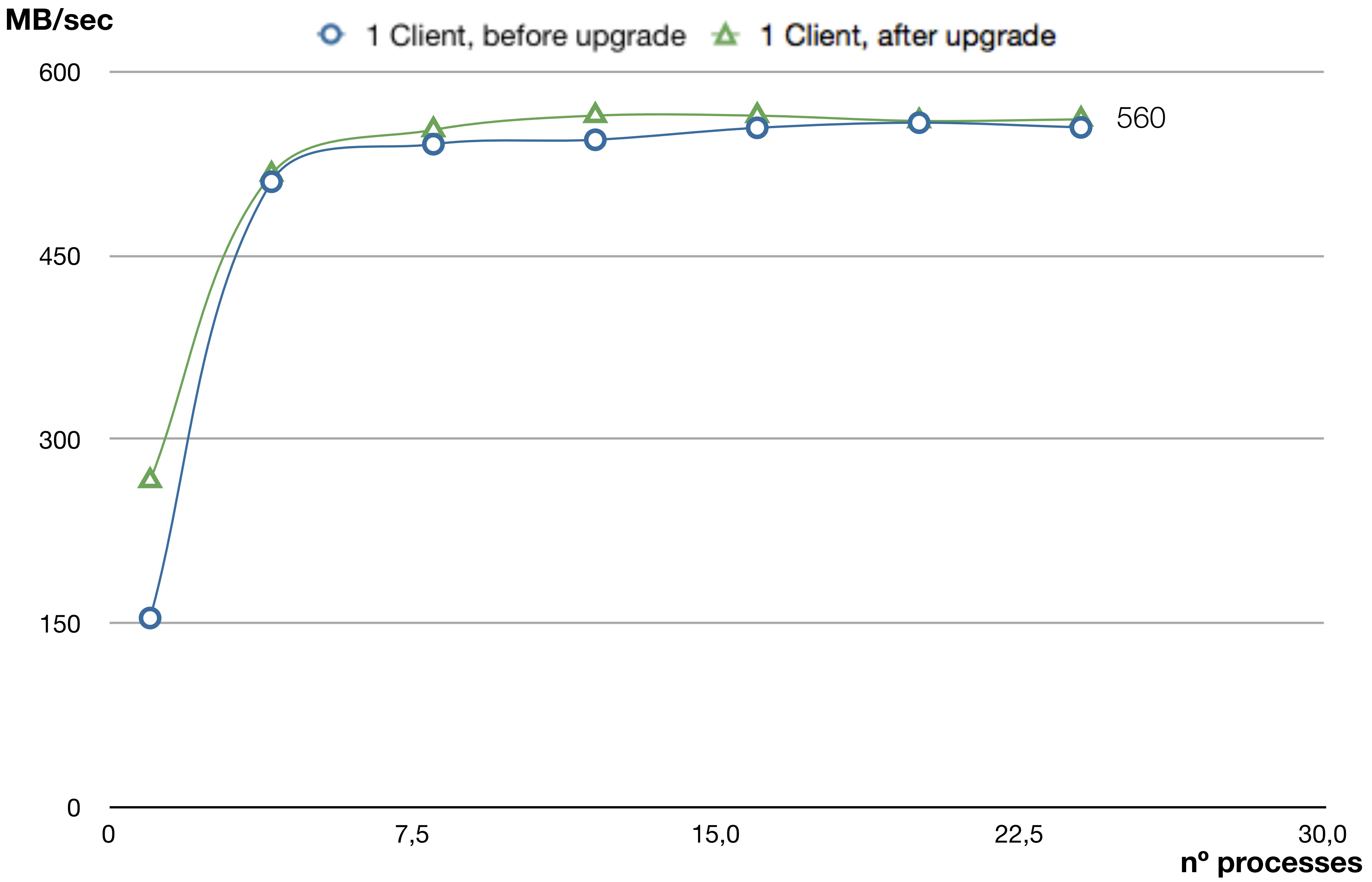
After upgrade slighly better, and both reach the bandwidth limit

100MB Upload


○ 1 Client, before upgrade
△ 1 Client, after upgrade
□ 5 Clients, after upgrade
MB/sec
600
450
300
150
0
0
25
50
75
100
nº processes
588,39
122

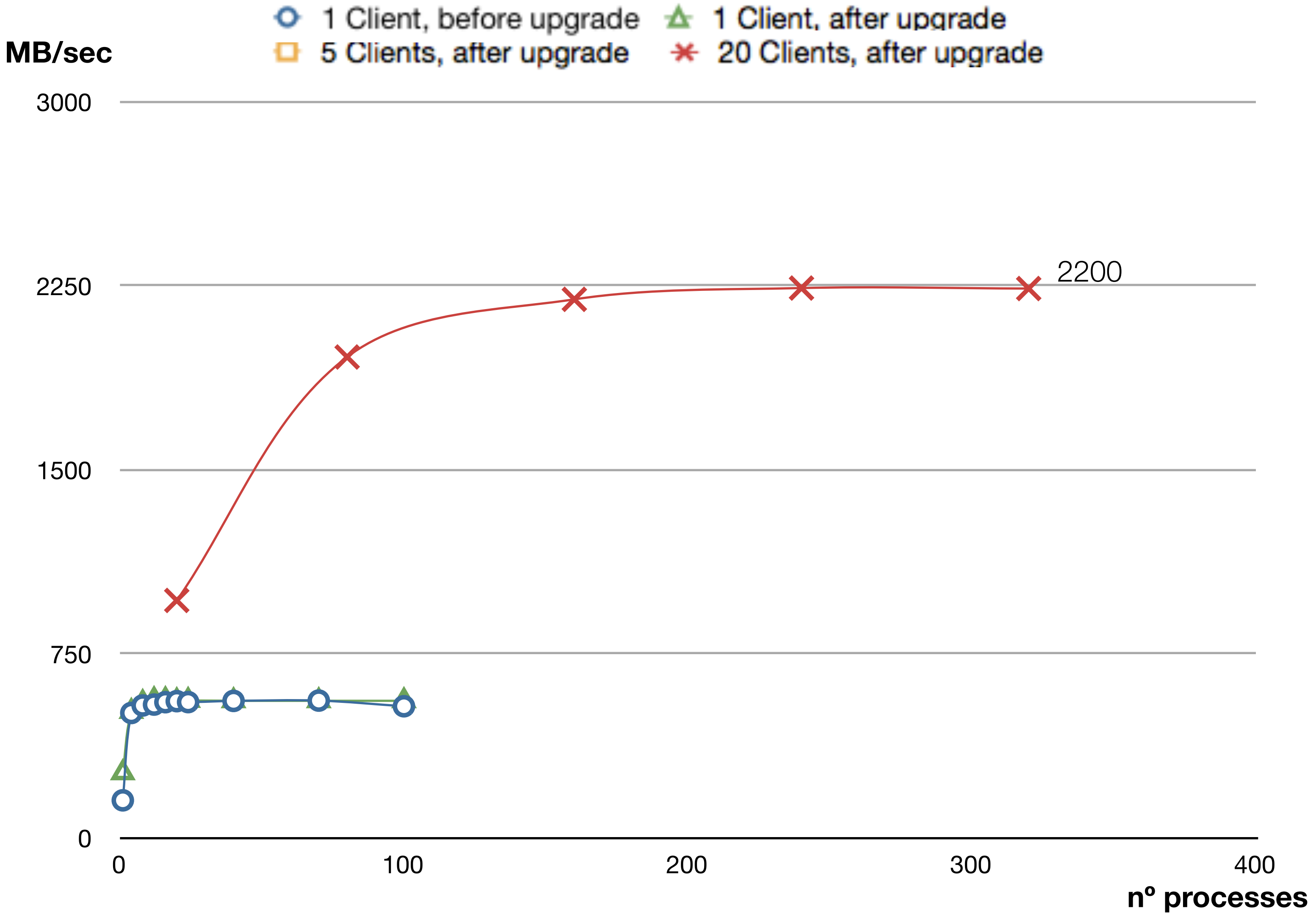
5 clients also reach bandwidth limit of 5Gb

100MB Download



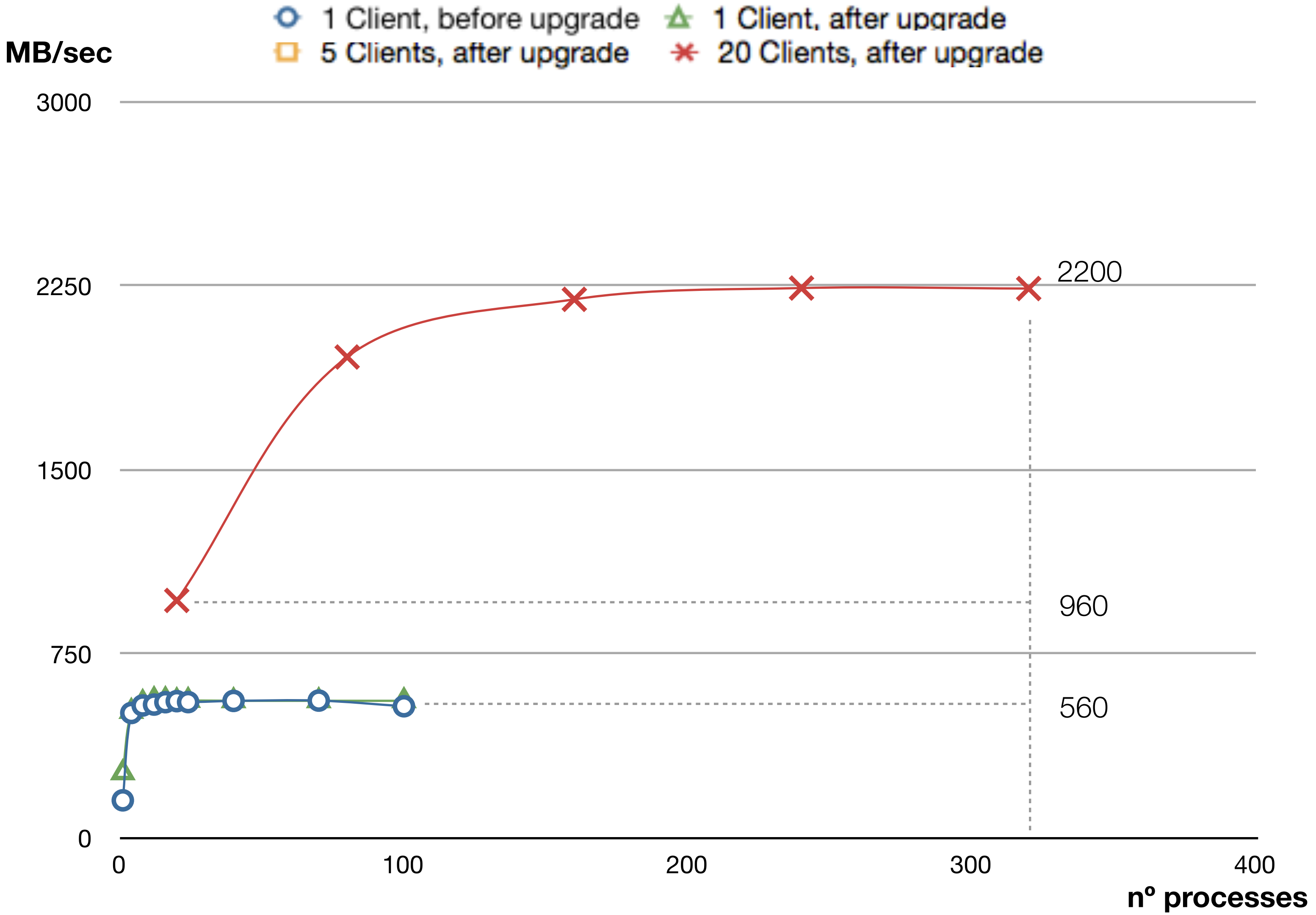
After upgrade slighly better, and both reach the bandwidth limit

100MB Download



20 clients reach bandwidth limit of 18Gb

100MB Download


○ 1 Client, before upgrade
△ 1 Client, after upgrade
□ 5 Clients, after upgrade
✱ 20 Clients, after upgrade
MB/sec
3000
2250
1500
750
0
0
100
200
300
400
nº processes
2200
960
560

20 clients reach bandwidth limit of 18Gb

Conclusion after testing phase

- Tested metadata performance up to 8k files/s
  - proved expected scalability of the system
- Tested total throughput up to 18 Gb/s
  - fully maxed out the 2 fibres available
  - balanced system with 350MB/s per OSC
- Fully achieved expected performance
- Minor technical problems found and resolved rapidly
  - very productive collaboration with clear benefits for CERN and HUAWEI

Future plans

- 2012 - short term
  - Further increase scalability range with additional network and client resources
  - Analyse performance impact of write cache and journal
  - Collect feedback ATLAS workload management system
  - Excercise transparent upgrade procedure
- 2013 - next year
  - Multiple datacenter tests (eg with IHEP)
  - Erasure code impact on performance and space overhead
  - Prove transparent disk failure recovery with consumer drives
- Ultimate goal 2013
  - Prove TCO gains of the system as part of a production service

Huawei Cloud Storage

Maitane Zotes Resines, CERN IT

Openlab Major Review Meeting
27. September 2012
CERN, Geneva