User’s file access

Current status

- The problem is not limited to one particular file
- Whenever we have simultaneous transfers of the same file, performance degrades significantly
- Glasgow is also observing this?

Example

- 5 simultaneous transfers of the same file from lxplus

```
[arogovsk@lxplus792 ~]$ bash test_copy.sh 5
Copying root://ceph-gw2.gridpp.rl.ac.uk:1094/lhcb:user/lhcb/user/a/arogovsk/test_hotfile
[DONE]  after 353s
Copying root://ceph-gw2.gridpp.rl.ac.uk:1094/lhcb:user/lhcb/user/a/arogovsk/test_hotfile
[DONE]  after 354s
Copying root://ceph-gw2.gridpp.rl.ac.uk:1094/lhcb:user/lhcb/user/a/arogovsk/test_hotfile
[DONE]  after 356s
Copying root://ceph-gw2.gridpp.rl.ac.uk:1094/lhcb:user/lhcb/user/a/arogovsk/test_hotfile
[DONE]  after 355s
Copying root://ceph-gw2.gridpp.rl.ac.uk:1094/lhcb:user/lhcb/user/a/arogovsk/test_hotfile
[DONE]  after 350s
[arogovsk@lxplus792 ~]$
```

Example

- 5 simultaneous transfers of different replicas of the same file from lxplus

```
[arogovsk@lxplus792 ~]$ bash test_copy.sh 5
Copying root://ceph-gw2.gridpp.rl.ac.uk:1094/lhcb:user/lhcb/user/a/arogovsk/test_hotfile2
[DONE]  after 19s
Copying root://ceph-gw2.gridpp.rl.ac.uk:1094/lhcb:user/lhcb/user/a/arogovsk/test_hotfile5
[DONE]  after 23s
Copying root://ceph-gw2.gridpp.rl.ac.uk:1094/lhcb:user/lhcb/user/a/arogovsk/test_hotfile3
[DONE]  after 27s
Copying root://ceph-gw2.gridpp.rl.ac.uk:1094/lhcb:user/lhcb/user/a/arogovsk/test_hotfile4
[DONE]  after 30s
Copying root://ceph-gw2.gridpp.rl.ac.uk:1094/lhcb:user/lhcb/user/a/arogovsk/test_hotfile1
[DONE]  after 31s
[arogovsk@lxplus792 ~]$
```

Vector read: current status

In previous episodes…

- Vector read operation is slow on echo
- Vector read is executed as sequential reads with rados striper
- If we remove rados striper and read directly from ceph objects, performace seem to improve
- This does not work with xcache
  - A lot of strange errors “file name too long”, due to a bug in xcache
- “local” caching was combined with non-striper reads

New tests

A new test suite that runs LHCb WG-production jobs (and possibly other types of jobs) locally on the WN was created

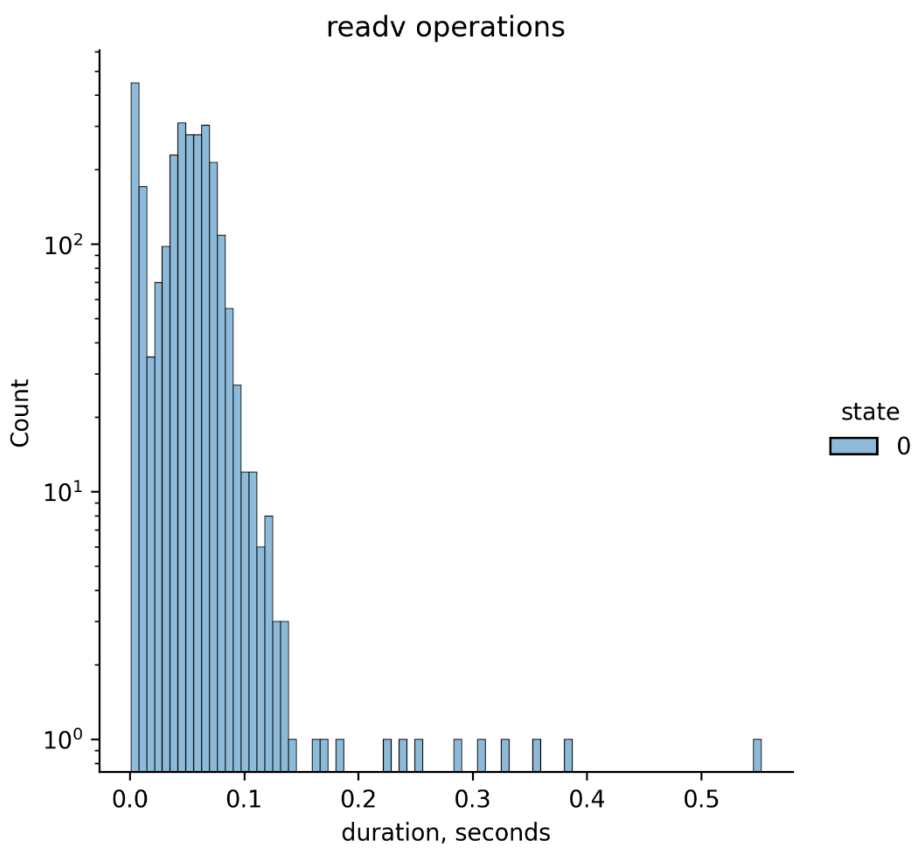
- jdls of the jobs are needed

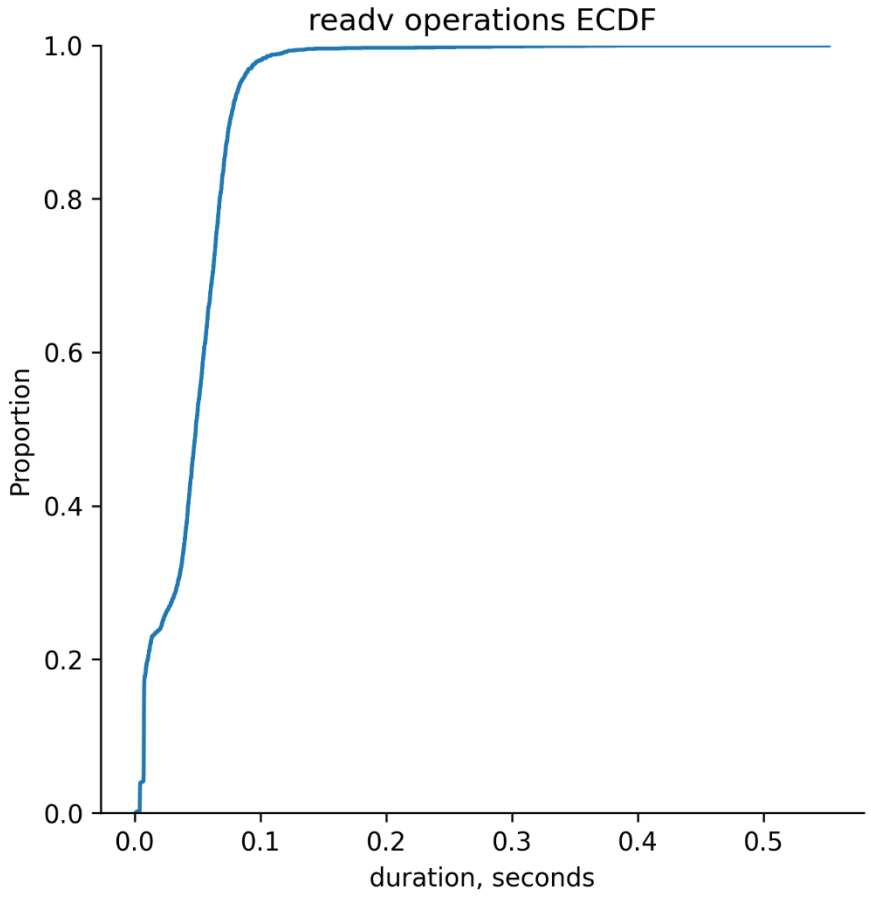
Test setup

A dedicated WN (lcg2270) was put out of production. Changes were applied to the gateway on the WN. Test setup is the following:

- 1 LHCb WG-production job

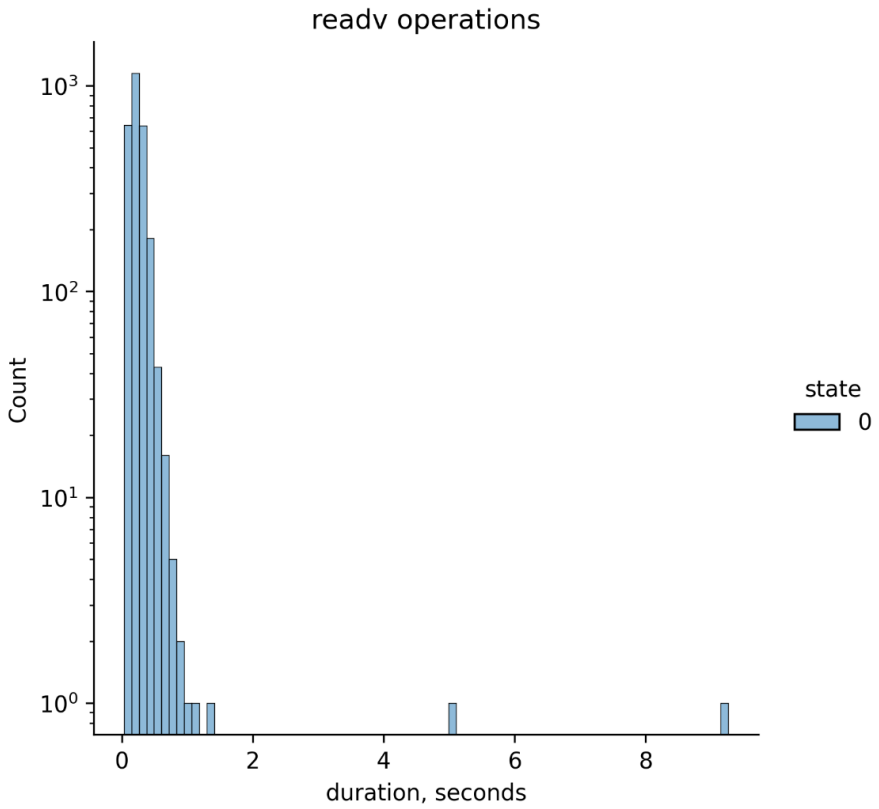
Results (readv, no changes)

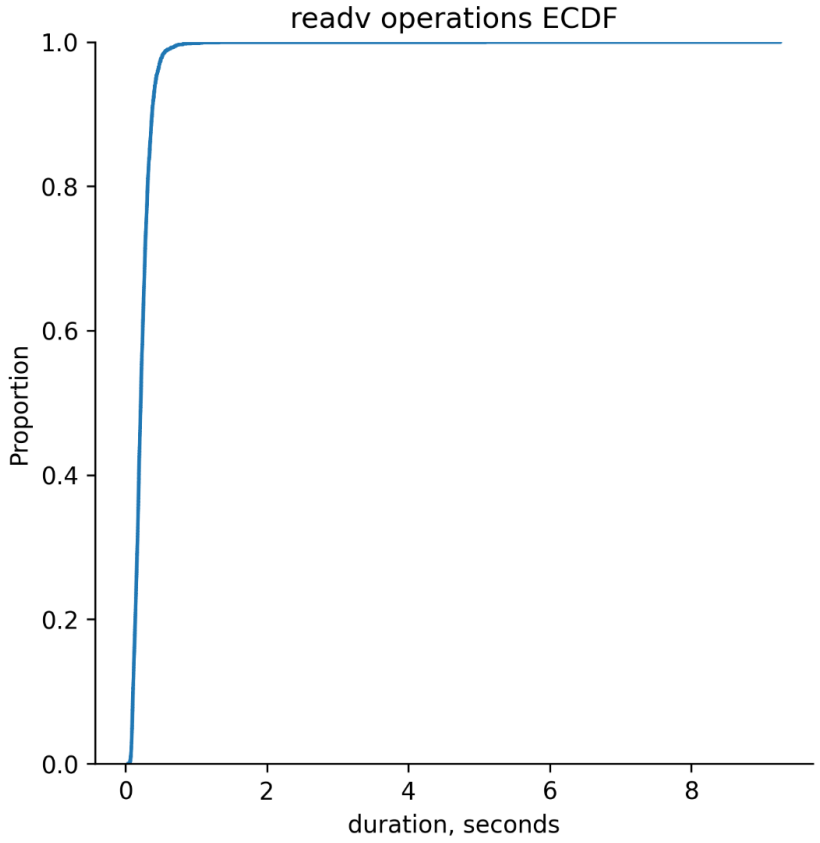




Median time: 0.049
Mean time = 0.047

Results (readv, patched readv,no xcache)





Median: 0.21,
Mean: 0.24

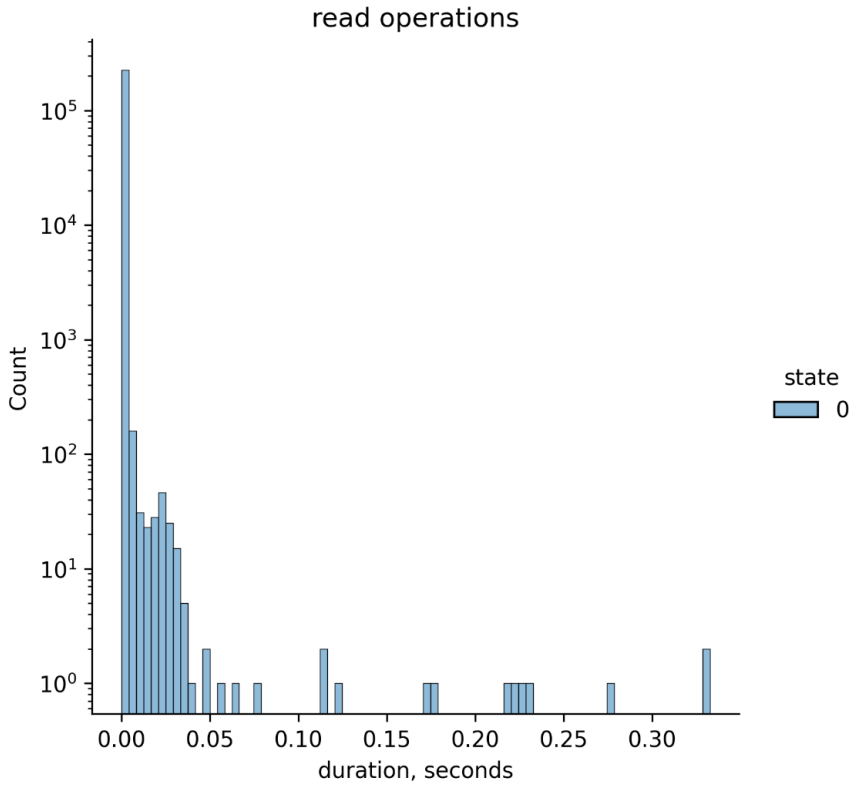
Results

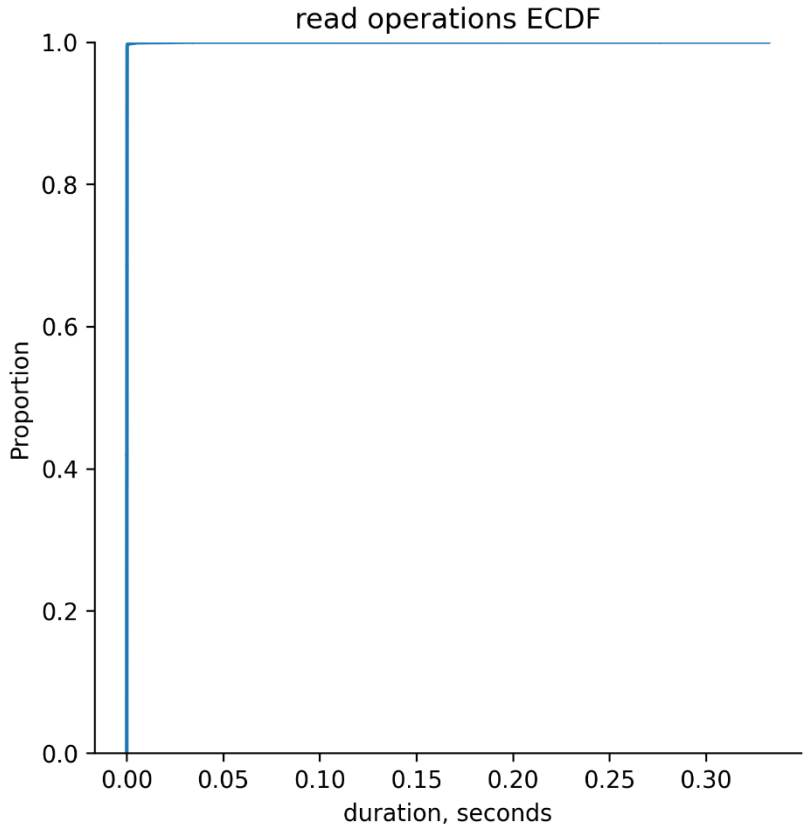
Below is the overall execution times for the jobs.

No changes	Patched readv
1 hour	4.5 hours

Why it is so slow with changes?

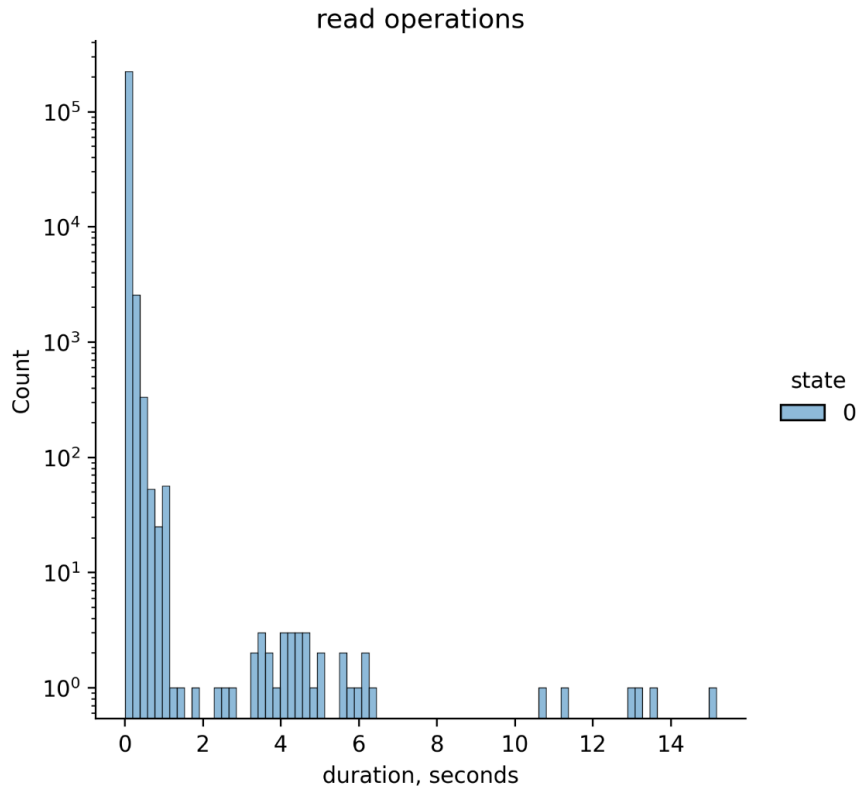
Results (reads, no changes)

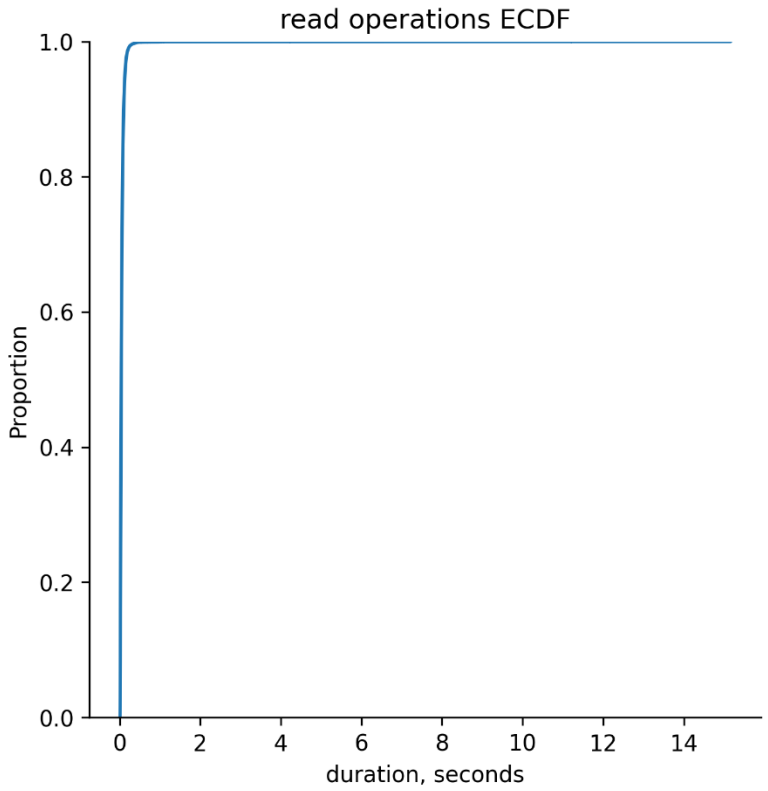




Median time: 0.0001
Mean time: 0.0001

Results (reads, readv patched, no xcache)



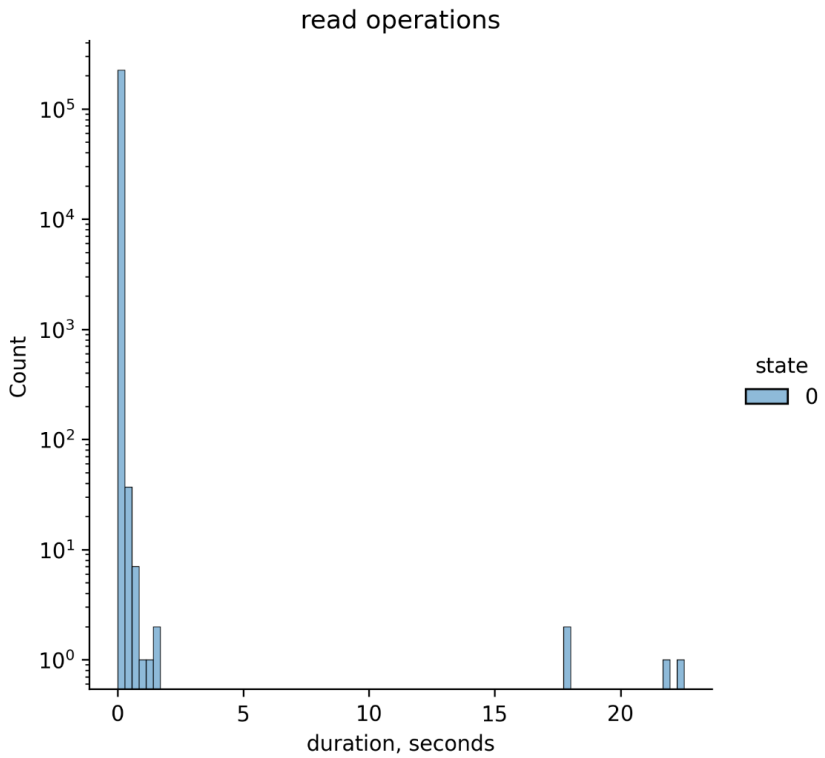


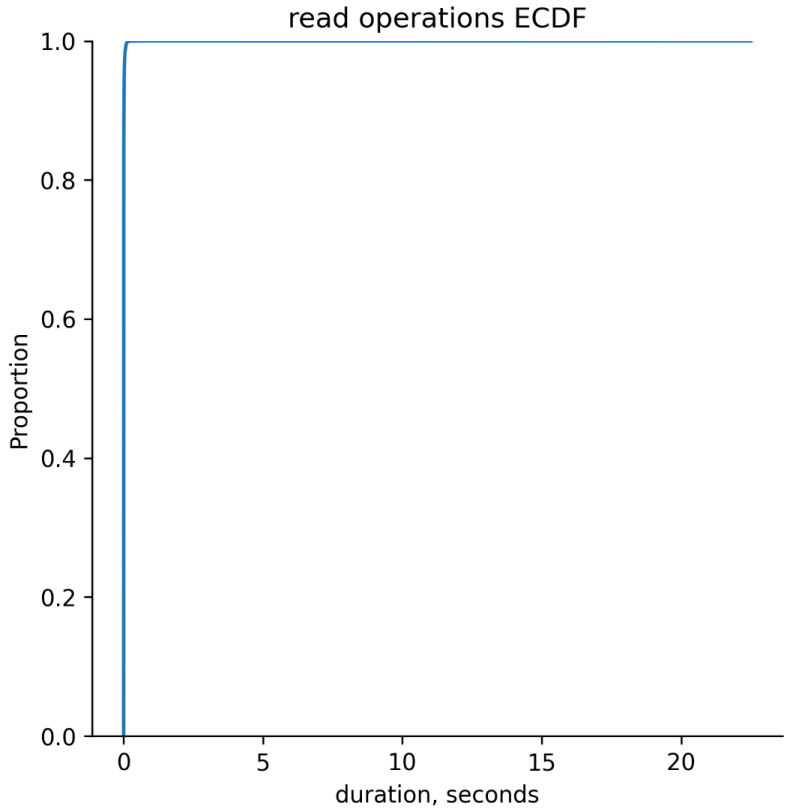
Median time: 0.04
Mean time: 0.05

Results

- Under small load current configuration performs better then the patched one, most probably due to xcache presence
- The difference is much greater for the read operations
- This is the reason to use non-striper operations for reads also

Results (reads, all patched, no xcache)





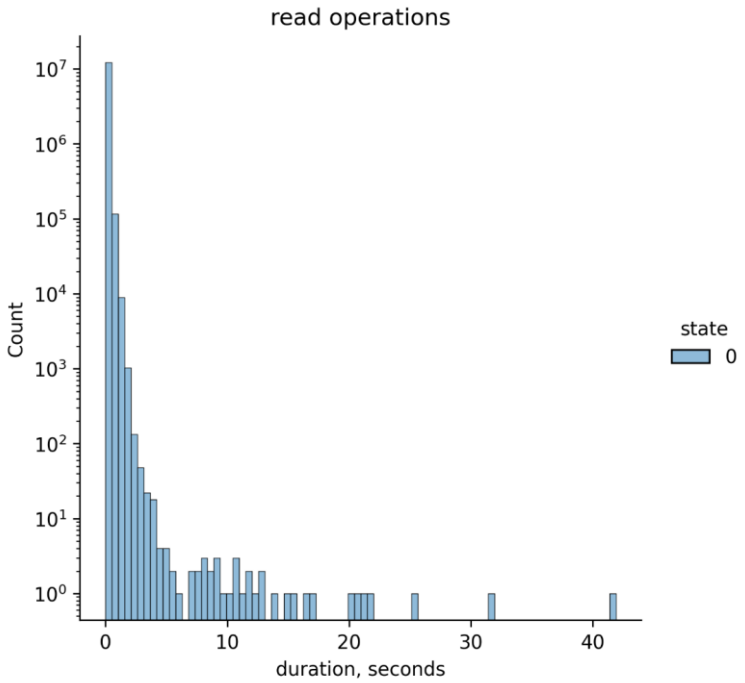
Median time: 0.003
Mean time: 0.006

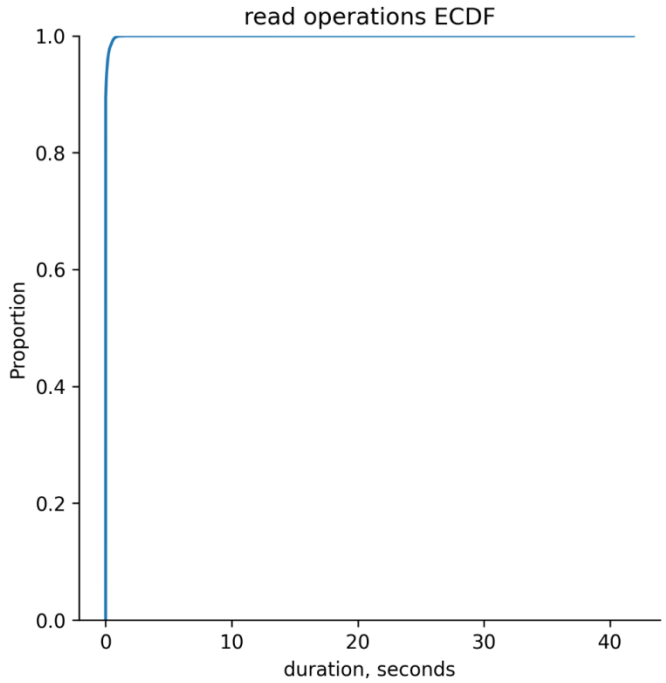
Test setup

A dedicated WN (lcg2270) was put out of production. Changes were applied to the gateway on the WN. Test setup is the following:

- 64 LHCb WG-production job

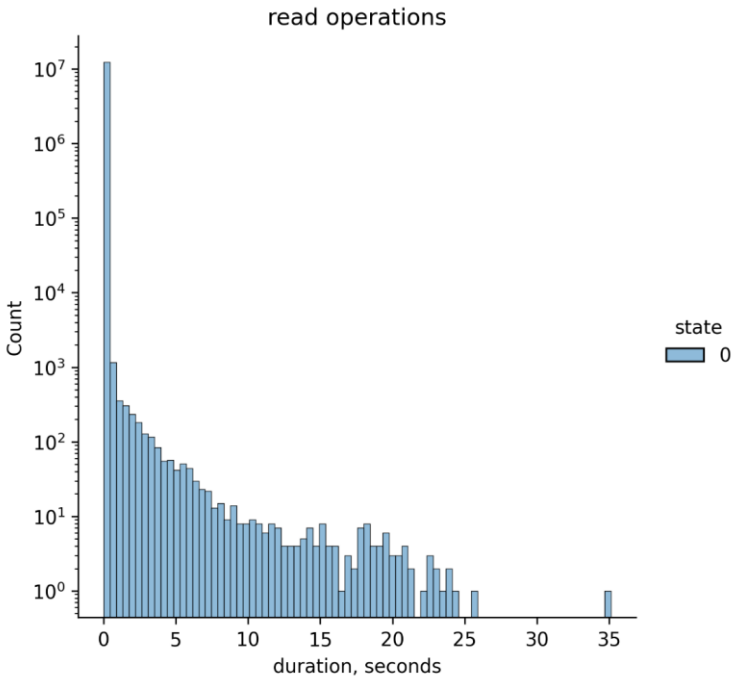
Results (reads, no changes)

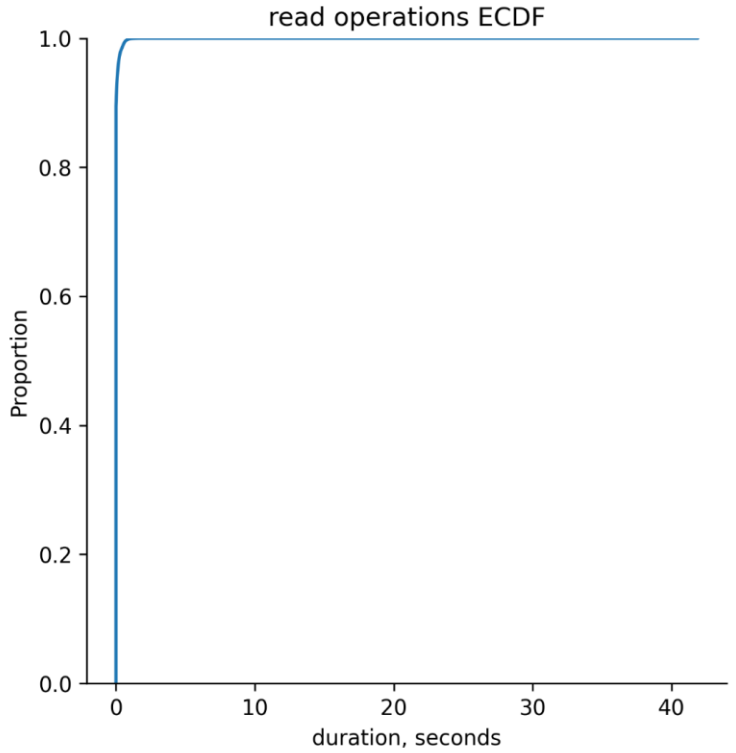




Median time: 0.0001
Mean time: 0.02

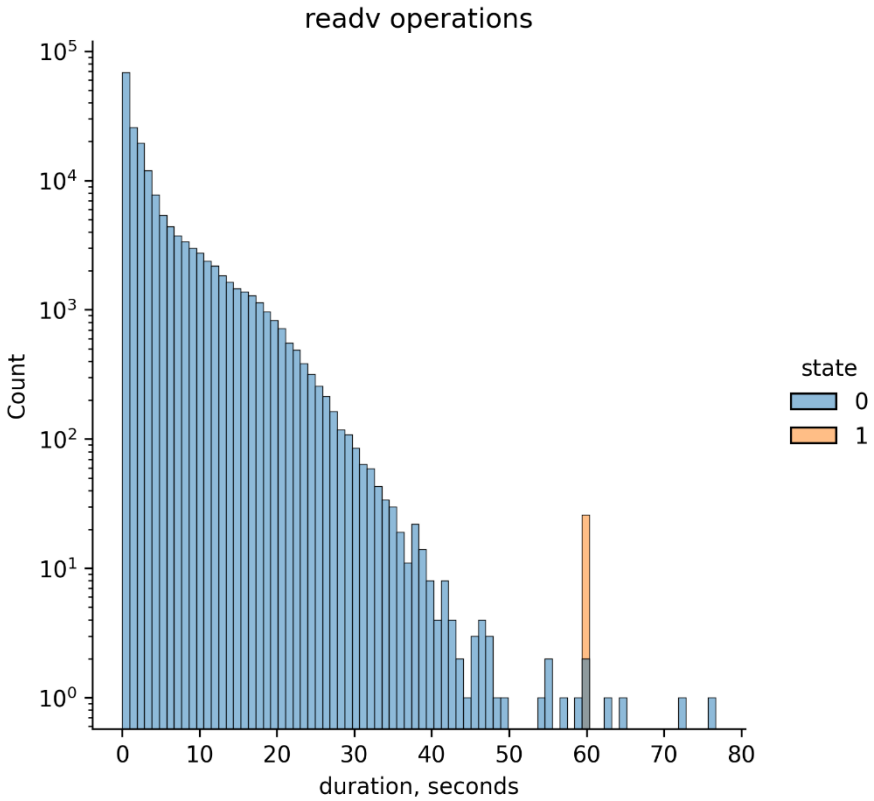
Results (reads, all patched)

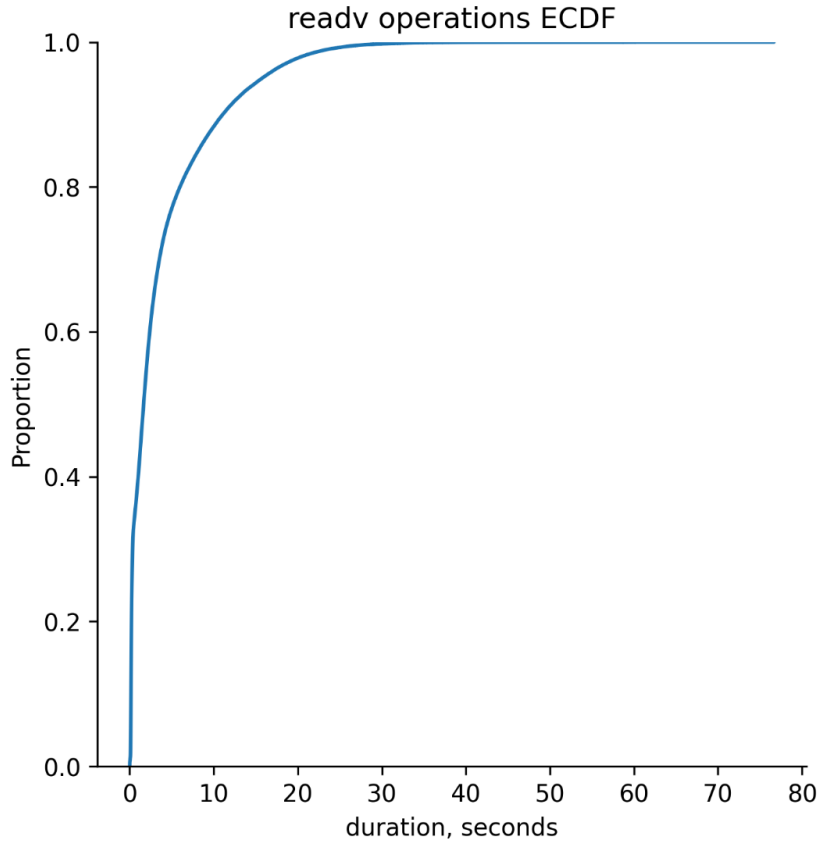




Median time: 0.003
Mean time: 0.009

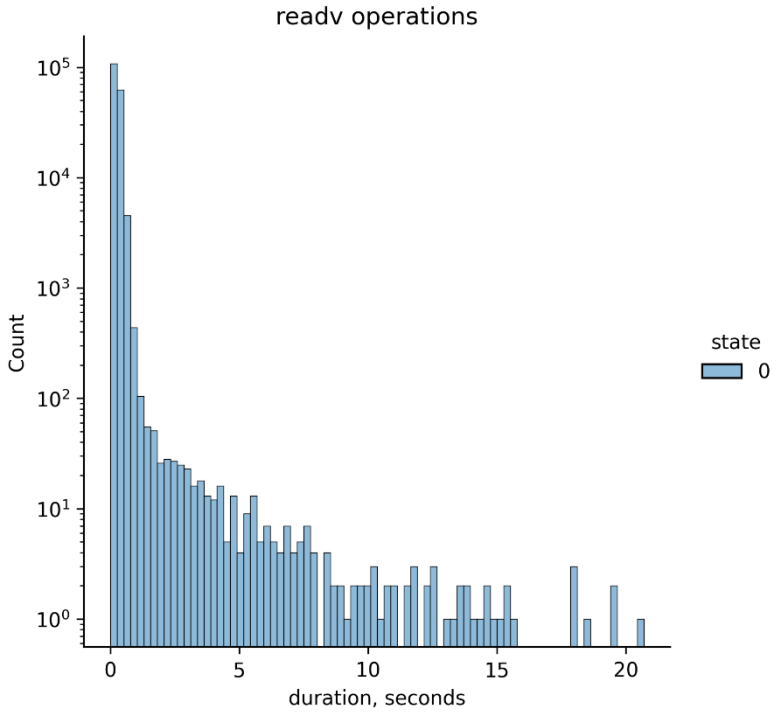
Results (readvs, no changes)

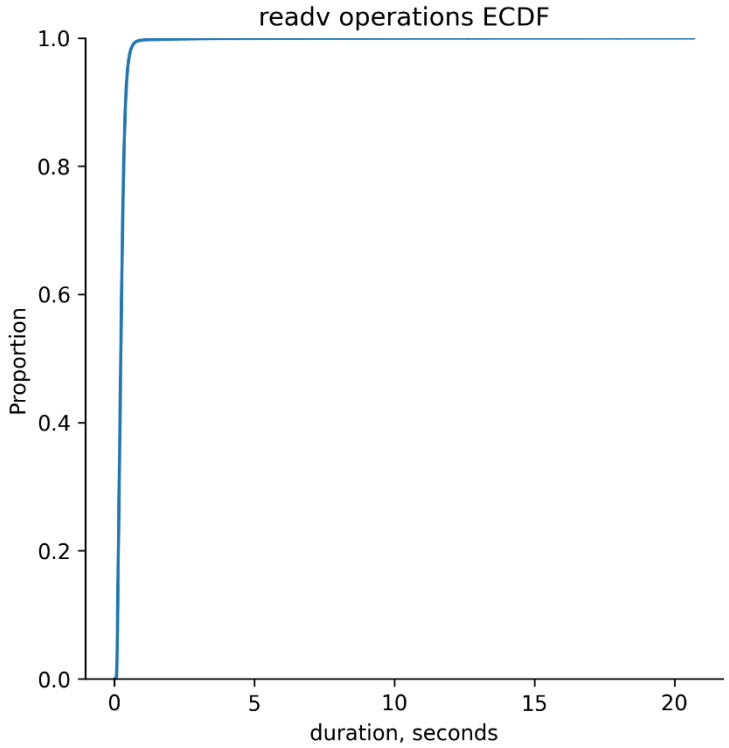




Median time: 1.65
mean time: 3.72

Results (readvs, all patched)





Median time: 0.23
Mean time: 0.26

Discussion

- readv:
  - Patched configuration look better for production jobs under significant load, and not so bad with small load
- read
  - Patched configuration is worse than the “standard” one, especially if reads are executed via striper. With significant load difference becomes smaller.
  - In case of efficiency drop on patched environment we may return to configuration with xcache..

To Do

- Tests in production
- Try mixing different types of jobs locally on WN?

Production tests

- Dedicated set of WNs will be assigned to LHCb-only jobs
- Some of the nodes (half) will be patched
- Should we preserve local gateway as read-only in patched environment?